BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

2881

Attorney Docket No. SURR.78

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANT:	HASTINGS		
SERIAL NO.:	09/994,575	EXAMINER:	KALIVODA, C.M.
FILED:	NOVEMBER 27, 2001	ART UNIT:	2881
TITLE:	MEDIAN FILTER FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA	CONF. NO.:	7303

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

REQUEST FOR RECONSIDERATION

Sir:

An Office Action was mailed in the above-captioned application on October 3, 2003. In such Office Action claims 1-5, 7-16, and 18-22 were pending. Independent Claims 1 and 12 were rejected under 35 U.S.C. § 102(e)(1). Dependent Claims 1 and 12 were rejected under 35 U.S.C. § 103(a). This Request for Reconsideration is submitted in response to such Office Action.

The Rejection under 35 U.S.C. § 102(e)(1)

Independent claims 1 and 12 stand rejected under 35 U.S.C. § 102(e) over a published patent application – Townsend et al., U.S. patent application Serial No. 09/950,313 (Pub. No. 2002/0102610) ("Townsend et al."). In rejecting the claims, the Examiner specifically cited paragraph 0057 (lines 1-6) of Townsend et al.

A published U.S. patent application is available as prior art under 35 U.S.C. § 102(e) as of its earliest effective U.S. filing date, "taking into consideration any proper benefit claims to prior U.S. applications under 35 U.S.C. 119(e) or 120 if the prior application(s) properly supports the subject matter used to make the rejection." Manual of Patent Examining Procedure (8th Ed., Rev. 1) (the "MPEP"), § 706.02(f)(1) (p. 700-27, 1st col.) (emphasis added).

37 CFR 1.8
CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on 3/23/04

Signature: Tasha L. Cove
Name: Tasha L. Cove

Townsend et al., filed on September 10, 2001, claims the benefit of U.S. provisional application Serial No. 60/232,273, filed on September 12, 2000 ("the '273 provisional"), a copy of which is attached hereto at Tab A. However, the '273 provisional does not contain the paragraph cited by the Examiner in rejecting the present application. Nor does it teach or suggest the use of a median filter. Because the '273 provisional does not support the subject matter used to make the rejection, the prior art date of Townsend et al. is no earlier than its actual filing date, September 10, 2001.

The present application was filed claiming the benefit of (i) U.S. provisional application Serial No. 60/253,178, filed November 27, 2000 ("the '178 application") and (ii) U.S. provisional application Serial No. 60/314,996, filed August 24, 2001 ("the '996 application"). The claimed invention is fully supported by these priority documents, copies of which are attached hereto at Tabs B and C, respectively.

Because the prior art date of Townsend et al. is antedated by the effective filing date of the present application, Townsend, et al. is not available as prior art. Reconsideration of the rejection pursuant to 35 U.S.C. § 102(e) is respectfully requested.

The Rejection under 35 U.S.C. § 103(a)

Dependent claims 4, 11, 15 and 22 stand rejected under 35 U.S.C. § 103(a) over Townsend et al. Dependent claims 2, 3, 5, 7-10, 13, 14, 16 and 18-21 stand rejected under 35 U.S.C. § 103(a) over Townsend et al. in view of McLafferty et al. (U.S. Patent No. 3,997,298).

In both cases, the Examiner's rejections depend on Townsend et al. being prior art under 35 U.S.C. § 120(e). As explained, *supra*, Townsend, et al. is not available as prior art against the present application because the effective filing date of the present application antedates the prior art date of the reference. Accordingly, Applicant requests that the rejection under 35 U.S.C. § 103(a) be withdrawn.

Closing Remarks

Applicant believes that the pending claims are in condition for allowance. If it would be helpful to obtain favorable consideration of this case, the Examiner is encouraged to call and discuss this case with the undersigned.

This constitutes a request for any needed extension of time and an authorization to charge all fees therefore to deposit account No. 19-5117, if not otherwise specifically requested. The undersigned hereby authorizes the charge of any fees created by the filing of this document or any deficiency of fees submitted herewith to be charged to deposit account No. 19-5117.

Respectfully submitted,

Date: March 23, 2004

Darla G. Yoerg, #48,053
Swanson & Bratschun, L.L.C.
1745 Shea Center Drive, Suite 330
Highlands Ranch, Colorado 80129
Telephone: (303) 268-0066
Facsimile: (303) 268-0065

S:\CLIENT FOLDERS\SURROMED\78\UTILITY\OFFICE ACTION RESPONSE.DOC

13/00

COVER SHEET FOR PROVISIONAL APPLICATION FOR PATENT

Assistant Commissioner for Patents
Box PROVISIONAL PATENT APPLICATION
Washington, DC 20231

Sir:

This is a request for filing a PROVISIONAL APPLICATION under 37 CFR 1.53(c).

Docket Number	9195-0045-888	Type a plus sign (+) inside this box →	+

INVENTOR(s) APPLICANT(s)			
LAST NAME	FIRST NAME	MIDDLE INITIAL	RESIDENCE (CITY AND EITHER STATE OR FOREIGN COUNTRY)
Townsend Robinson	Robert Andrew	R	St. Louis, Missouri Suskatoon, Canada

TITLE OF THE INVENTION (280 characters max)

METHODS FOR THE SEQUENCING, IDENTIFICATION AND CHARACTERIZATION OF PEPTIDES OR PROTEINS AND USES THEREOF

CORRESPONDENCE ADDRESS:

PENNIE & EDMONDS LLP
1155 Avenue of the Americas
New York, NY 10036-2711
(212) 790-9090

ENCLOSED APPLICATION PARTS (*check all that apply*)

☒ Specification	*Number of Pages*	52	☐ Small Entity Statement	
☒ Drawing(s)	*Number of Sheets*	16	☒ Other (specify)	Large Entity

METHOD OF PAYMENT (*check one*)

☐ A check or money order is enclosed to cover the Provisional filing fees.	ESTIMATED PROVISIONAL FILING FEE AMOUNT	☒ $150 ☐ $75
☒ The Commissioner is hereby authorized to charge the required filing fee to Deposit Account Number 16-1150.		

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No. ☐ Yes, the name of the U.S. Government agency and the Government contract number are: _____

Respectfully submitted, [signature] (Reg. No. 36,581)

Signature Laura A. Coruzzi

Laura A. Coruzzi, Esq.
PENNIE & EDMONDS LLP

REGISTRATION NO. (*if appropriate*) 30.742

Date September 13, 2000

☐ Additional inventors are being named on separately numbered sheets attached hereto.

Total number of cover sheet pages. 1

PROVISIONAL APPLICATION FILING ONLY

METHODS FOR THE SEQUENCING, IDENTIFICATION AND CHARACTERIZATION OF PEPTIDES OR PROTEINS AND USES THEREOF

1. BACKGROUND OF THE INVENTION

It is generally recognised that a minor portion (<5%) of DNA sequences composing the genomes of higher plants and animals is used as a template for the biosynthesis of proteins. For example, of the approximately 3,000,000,000 base pairs in the human genome, an estimated 3% are translated into protein (Liang, et al. (2000) Nature Genetics 25, 239). The genomic segments which are expressed are contained within the exons which are separated by untranslated regions, or introns.(Gardiner, K (1996) Trends in Genetics 12, 519-524). It is this minor portion of 'functional' DNA that comprises the estimated 50,000 expressed human genes which form the basis of normal and disease phenotypes in man. This is because the information which identifies the translated segments of the human genome is not present within the nucleotide sequences alone. To unambiguously establish gene structure and to identify the regions which are translated to produce proteins, the gene products, mRNA and protein, must be analysed and linked to the parent genomic sequences (Martin, K.J. and Pardee, A.B. (2000) Proc. Natl. Acad. Sci. 97, 3789-3791, the contents of which is hereby incorporated by reference in its entirety).

Unfortunately, identification of expressed genes is only the first step toward understanding the role of proteins in normal and pathological cellular processes. Once translated, proteins become enzymatically modified at precise amino acid residues with specific groups, chemical changes in amino acid side chains can occur (e.g. oxidations and deamination) and often the active form of a gene product consists of specific proteolytic derivatives of the initial translated product (protein isoforms). Over 250 post-translational modifications have been described (See http://pir.georgetown.edu/cgi-bin/pirwww , the contents of which is hereby incorporated by reference in its entirety). An increasing number of proteins are being described in which internal sequences (inteins of protein introns) have been proteolytically removed to produce the active protein form (Mills, K.V., Lew, B.M. (1998) Proc. Natl. Sci. 95, 3543-8, the contents of which is hereby incorporated in its entirety). These modifications have profound effects on protein activity and macromolecular interactions in the living cell (Parekh, R.B. and Rohlff, C. (1997) Curr. Opin. Biotech. 8, 718-723, the contents of which is hereby

incorporated by reference in its entirety) and the modified, active protein forms can become the key targets for novel therapeutics.

Characterisation of the complement of expressed proteins from a single genome is a central focus of the evolving field of proteomics. Since one genome produces many proteomes (hundreds in multi-cellular organisms) and the number of expressed genes in a cell is minimally 10,000, the characterisation of thousands of proteins to evaluate proteomes can only be accomplished using a high-throughput, automated process. The sequencing of peptides using mass spectrometry (MS) is one approach that offers significant promise for the rapid sequencing of peptides required for industrialized proteomics.

The analysis of peptides using mass spectrometry began with the introduction and development of fast atom bombardment in the 1980's. Additional new 'soft' ionisation methods enabled the formation of ions from larger molecular weight biomolecules (500-200,000). The co-development of soft ionisation methods and tandem mass spectrometry produced an improved method for efficiently obtaining peptide fragmentation spectra of sufficient quality and information content for peptide sequencing. The significant advantages over the traditional protein sequencing method of Edman degradation are speed and the potential ability to characterise all amino acid residue modifications. Other methods which have been used to produce fragmentation spectra include post-source decay after matrix-assisted laser desorption ionisation with time-of-flight or Fourier transform mass analyzers and low energy collision-induced dissociation in ion-trap mass spectrometers. Instrumentation configured with two mass analyzers offer good performance with regard to sequence information in fragmentation spectra, resolution and mass accuracy of product ions.

In a tandem mass spectrometry experiment, a peptide sample is introduced into the mass spectrometer and is subjected to analysis in two mass analyzers (denoted as MS1 and MS2). In MS1, a narrow mass-to-charge window (typically 2Da), centered around the mass of the peptide to be analyzed, is selected. The peptide precursor ion is then subjected to fragmentation via collision-induced dissociation, which typically occurs in a collision cell by applying a voltage to the cell and introducing a gas to promote fragmentation. The process produces smaller peptide fragments derived from the precursor ion (termed the 'product or daughter ions'). The product ions, in addition to any remaining intact precursor ions, are then passed through to a second mass

spectrometer (MS2) and detected to produce a fragmentation (or MS/MS) spectrum. The MS/MS spectrum (which is commonly expressed with the mass-to–charge ratio on the horizontal axis and the intensity on the vertical axis) comprises peaks corresponding to the mass-to-charge ratios of all detected product and precursor ions as a function of their measured intensities. FIG. 1 shows a subset of fragment ions most useful for determining peptide sequence (nomenclature as described by Biemann). (K. Biemann, Biomed. Environ. Mass Spectrom. 16). Fragmentation across the peptide bond produces either a charged C-terminal fragment ion (y ions) or a complementary, charged N-terminal fragment ion (b ions). The a ions are related to the b ions by a mass decrement of 27.997 Da.

It is well recognised that fragmentation spectra of known peptides produced by tandem mass spectrometry contain a wealth of structural information which is consistent with the known sequence. Many investigations have focused on rationalising the observed fragmentation spectra with the chemical structures of the known peptides. These studies involved manually interpreting spectra and focused on formulating mechanisms of peptide gas-phase chemistry. The scientific rationale of these studies (encapsulating a large research effort in the 1980's in biomolecular mass spectrometry) was that a set of rules would emerge for the *ab initio* determination of the correct peptide sequence from fragmentation spectra. Although the result was a significant increase in the understanding of peptide fragmentation mass spectrometers and a comprehensive cataloguing of fragmentation ions, it became apparent that the tandem spectrum from a single known peptide could be interpreted to be consistent with multiple sequences. This lack of a unique result is a major impediment to the development of accurate, high throughput methods for sequencing peptides using tandem mass spectrometry. Therefore, although peptide sequencing by MS/MS has the advantage of very short data generation time, without considerable advances in data interpretation this benefit is difficult to exploit.

The ambiguities associated with spectral interpretation spawned various computer algorithms for limiting the number of possible interpretations of MS data and thus, deducing the sequence of a peptide from an MS/MS spectrum. This effort has continued from the 1980's. (Hamm et al. CABIOS 2, 115-118 (1986); Sakurai et al. Biomed. Mass Spectrom. 11, 396-399 (1984), the contents of which are hereby incorporated by reference in their entirety). One approach involved the use of software

to evaluate all possible amino acid sequences corresponding to the molecular mass of a peptide. This was problematic in that it produced too many candidate sequences for the correct sequence to be identified with certainty. For example, a peptide of nominal mass 1000 Da can yield more than 10^{10} possible amino acid sequences if the 20 naturally occurring amino acids are considered in calculating all sequence permutations. Algorithms were therefore developed based on 'sub-sequencing' strategies whereby portions of the total sequence, or sub-sequences, were tested against the mass spectrum. (Ishikawa et al. Biomed. Environ. Mass Spectrom. 13, 373-380 (1986); Siegel et al. Biomed. Environ. Mass Spectrom. 15, 333-343 (1988); Johnson et al. Biomed. Environ. Mass Spectrom. 18, 945-957 (1989); Yates III, P. R. Griffin and L. E. Hood, in Techniques in Protein Chemistry, edited by J. J. Villafranca, Vol. 2, Academic Press, San Diego pp.. 477-485 (1991)), the contents of which are hereby incorporated by reference in their entirety). The sub-sequences that correlate to ions observed in the MS/MS spectrum are extended by a residue and the whole process is then repeated until the entire sequence is obtained. During each incremental extension of the sequence, the possibilities are reduced by comparing sub-sequences with the mass spectrum and only permitting continuation of the process for sub-sequences giving the most favourable spectral matches. Determination of amino acid composition (either by chemical analysis or accurate mass measurement of ions in the low mass-to-charge region in the spectrum corresponding to amino acid constituents) has also been utilised to limit sequence possibilities. (Zidarov et al. Biomed, Environ. Mass Spectrom. 19, 13-16 (1990), the contents of which is hereby incorporated by reference in its entirety). One of the difficulties of these 'sub-sequencing' methods is that, in cases where particular fragment ions are absent or of low intensity in a single spectrum, the correct sequences may fail to be considered.

An alternative approach has been to develop programs for *de novo* peptide sequencing from fragmentation spectra based on a graph theory approach. (Fernandez-de-Cossjo, J. Gonzalez and V. Besada, CABIOS 11, 427-434 (1995); Hines, A. M. Flick, A. L. Burlingame and B. W. Gibson, J. Am. Soc. Mass Spectrom. 3, 326-336; Knapp, J. Am. Soc. Mass Spectrom. 6, 947-961 (1995)), the contents of which are hereby incorporated by reference in their entirety). The basic method involves mathematically transforming an MS/MS spectrum to a form where fragment ions are converted to a single fragment ion type represented by a vertex on the spectrum graph. (Bartels, Biomed. Environ. Mass

Spectrom. 19, 363-368 (1990)), the contents of which is hereby incorporated by reference in its entirety). Peptide sequences are then determined by finding the longest series of these transformed ions with mass differences corresponding to the mass of an amino acid. However, these types of programs have the potential to produce numerous matches for a given peptide sequence. Furthermore, errors in sequence determination can occur where sequence ions are absent or not of sufficient intensity to be detected or where noise spikes in the spectrum are falsely assigned as fragment ions. In summary, although it was recognised that the interpretation of peptide fragmentation spectra from tandem mass spectrometry might hold significant promise for sequencing unknown peptides, *ab initio*, interpretation of a known peptide by the existing methods tend to give multiple sequences, of which the correct one is difficult to unequivocally identify.

Another significant conceptual advance for determining peptide sequences from fragmentation spectra occurred with methods which attempt to match spectral information with sequences in protein and translated nucleotide sequence databases. (Henzel et al. Proc. Natl. Acad. Sci. U.S.A. 90, 5011 (1993); Eng et al., J. Am. Soc. Mass Spectrom. 5, 976-989 (1994); Mann and M. Wilm, Anal. Chem. 66, 4390-4399 (1994); Clauser, P. Baker and A. L. Burlingame, in Proceedings of the 44th ASMS Conference of Mass Spectrometry and Allied Topics. Portland, OR, 1996, pp.365-366), the contents of which are hereby incorporated by reference in their entirety). Although this approach reduces the number of peptide sequence possibilities and a correct match identifies the entire protein or nucleotide sequence in the databases, multiple, the results equally probable 'hits' are often. One error-tolerant algorithm for searching protein and nucleotide databases with mass and sequence information from fragmentation spectra of tryptic peptides has been described (MS-TAG) (Mann and Wilm (1994) Anal Chem. 66, 4390). Partial amino acid sequences are manually read from fragmentation spectra. The partial sequences are determined from the differences between a sequential set of m/z values in the spectrum and the masses of the 20 naturally-occurring amino acids. No constraints on the length or amino acid composition of the partial sequences are described. The molecular weight of the peptide, the ion read (y ion versus b ion) and the partial sequence are entered into the program. A search string of the general form, mass1-sequence-mass2, is calculated from the N and C-terminal mass deficits from the start and end of the selected partial sequence. No constraints on the value of the flanking masses is described. Multiple mass-sequence-mass constructs are used to iteratively

search a database. For each peptide sequence within a protein retrieved from the database search, the theoretical suite of sequencing ions is calculated and compared to the experimental fragmentation spectrum. From the manual comparison between the tandem spectrum and the database sequences, a set of protein or conceptually translated nucleotide sequences which share the same amino acid sequence within the confines of the mass-sequence-mass search string is chosen. Limitations of the MS-TAG approach are that overall throughput is limited by producing partial sequences using manual spectral interpretation and the well-recognised problem that sequences can be determined incorrectly (Perkins et al. (1999) Electrophoresis 20, 3551-3567), the contents of which is hereby incorporated by reference in its entirety). Thus, a preferred implementation of the method (to aid spectral interpretation) involves obtaining fragmentation spectra from the same peptide after derivatization of the carboxyl groups by methylation or incorporating ^{18}O into the C-terminal carboxy group during enzymatic digestion is recommended (Schevchenko et al. J. of Protein Chemistry and Schevchenki, A. (1997) Rapid Commun. In Mass Spectr. 11, 1015-1024, the contents of which is hereby incorporated by reference in its entirety). A similar approach has been extended to the analysis of intact proteins using laser fragmentation and Fourier-transform mass spectrometry in which four proteins were successfully identified (Mortz, E. et al. (1996) PNAS 93, 8264, the contents of which is hereby incorporated by reference in its entirety).

A different approach, which identifies database peptide sequences by comparing the experimental fragmentation spectrum with theoretical spectra from a mass-constrained set of database sequences has been described (SEQUEST) (Yates III et al. in U.S Patent No. 5,538,897 the contents of which is hereby incorporated in entirety into the present application). For each candidate database spectrum, a theoretical fragment spectrum is formed according to a selected ion model of peptide fragmentation. The predicted mass spectra are compared to the experimentally derived fragment spectrum by a cross-correlation function for scoring spectra. Although the list of possible sequences is significantly reduced and ranking schemes are described to produce the 'best fit' sequences from a database, identifying the correct protein or translated nucleotide sequence nevertheless still requires a manual assessment of the experimental spectrum and the sequences returned from the database.

A third approach uses partial amino acid sequences to increase the specificity of mass matching algorithms. This approach typically involves applying the interpreted sequence constraints to the list of database entries that result from searching an *in silico* tryptic digest of a protein or translated nucleotide database. The sequences may be a result of either Edman degradation, post-source decay induced by laser desorption ionisation or from tandem mass spectrometry. A mass matching algorithm has been described which incorporates sequence information with mass matching results into a probability scoring method and the results verify the well-known observation that partial sequence information in combination with mass matching increases the probability of finding the correct database hit. It should be emphasised that the use of the partial sequence data differs significantly from the mass-sequence-mass approach described above. In the mass-matching case, the partial sequence is not linked to flanking masses within the confines of a peptide defined by the specificity of a specific endoprotease. All three are susceptible to the major impediment in converting peptide fragmentation spectra into the correct database sequences—false positives. Thus, an automatable process has not been described which can accurately match peptide fragmentation spectra with the correct expressed sequences in protein or nucleotide databases.

In one embodiment, the present invention describes a user-independent method to translate mass spectrometric data from protein-derived peptides into a set of genome database sequences which share a single, correct peptide sequence or sequences. In a preferred embodiment, the method uses solely computer algorithms and thus, bypasses the subjective, time-consuming constraints associated with manual interpretation and conversion of fragmentation mass spectra to candidate sequences, produced by either *de novo* or database-assisted approaches. The retrieval of singular template sequences from complex, polycistronic genomic databases enables the high-throughput identification and organization of expressed segments of genomic DNA and the rapid identification of natural and artefactual nucleotide sequencing errors.

2. SUMMARY OF THE INVENTION

In one embodiment, the present invention describes a method for defining the physical DNA sequences in genomes which are expressed as proteins. The accuracy and robustness of the process is sufficient for use with large polycistronic genomes which contain the inherent errors of state-of the-art nucleotide sequencing methods and maybe

unedited and unassembled. In one embodiment, the invention consists of linked software modules to create an automated flow of data from primary mass spectral information to expressed gene sequences. The Holistic Peptide Sequencing (HOPS) module uses a novel spectral sequencing algorithm to produce sequence information with high accuracy from peptide fragmentation mass spectra. In one embodiment the Find Related Peptide (FIREPEP) software constructs search strings from selected HOPS sequences and the molecular weights of the respective peptides, and uses database specific criteria for allowed partial sequences to construct the database 'hook' string. The 'hook' string is then used to search a database of protein sequences and/or conceptually translated genome sequences, and/or peptide sequences. The result of the search is a single set of raw translated genome sequences and/or protein sequences which share the same HOPS peptide sequences found in the 'hook' search string. In one embodiment the Find Related Protein (FIREPROT) module consolidates the translated genome sequences and/or protein sequences by mapping all remaining HOPS sequences and observed peptide molecular weights onto the identified sequences. This is a description of an user-independent, high fidelity method which uses protein-derived mass spectrometric data (fragmentation spectra and measured peptide molecular weights) to identify unambiguously protein sequences and/or expressed regions in genome sequences. This method may be used to organize intron and exon sequences into a coherent gene structure, correct artifactual nucleotide sequencing errors, define expressed genetic mutations and protein polymorphisms, characterise post-translational proteolytic processing and define the type, and the residue location of amino acid residue modifications.

3. BRIEF DESCRIPTION OF THE FIGURES

FIG. **1** shows the suite of MS/MS peptide sequence ions required for the present invention (nomenclature of Biemann).

FIG. **2** summarises an embodiment the process for identifying protein sequences in genomes and assigning all sequence and mass data to additional regions of expressed sequence or to post-translational modifications.

FIG. 3 shows the modules in the HOPS algorithm (Holistic Peptide Sequencing) for the interpretation of peptide fragmentation spectra according to a preferred embodiment of the present invention

FIG. 4 details an embodiment of the steps in HOPS for constructing and editing the peak table derived from the fragmentation mass spectrum.

FIG. 5 diagrams an embodiment of the main peptide sequencing module of HOPS.

FIG. 6 is a flow chart of one embodiment for editing candidate peptide sequences from the sequencing module of HOPS.

FIG. 7 shows an embodiment of the steps involved in the ranking of sequences by number of ions matched and ion intensity in the peak list from the peptide fragmentation spectrum.

FIG. 8 details one embodiment of the steps involved in forming a candidate search string set from the top ranked HOPS sequences.

FIG. 9 shows a preferred embodiment of the (FIREPEP, Find Related Peptides) algorithm for searching an *in silico* tryptic digest with the database-specific 'hook' search string. (See Fig. 10 for search string structure for human genome).

FIG. 10 details an embodiment of the 'hook' string construction rules and the allowed criteria of the retrieved sequences from the six-frame translated human genome.

FIG. 11 shows an embodiment of the algorithm (FIREPROT, Find Related Proteins) to map all MS sequence data onto retrieved protein or translated nucleotide sequences.

FIG. 12 is an example of consolidating all MS data onto a gene sequence retrieved from the human genome and alignment of related cDNA and protein sequences.

FIG. 13 shows an embodiment of the algorithms to map observed peptide masses and post-translational modifications onto the unique set of identified translated genome sequences.

FIG. 14 shows the mapping of observed masses and protein phosphorylation onto translated genome sequences.

4. DETAILED DESCRIPTION OF THE INVENTION

In one embodiment of the present invention, a method is described for interpreting fragmentation mass spectra of modified and unmodified polypeptides to obtain sequence information with high accuracy and fidelity. The method is implemented as an automated procedure, that uses an algorithm we have named HOPS (Holistic Protein Sequencing), for interpretation of a peptide fragmentation mass spectrum without reference to any pre-determined peptide or protein sequence or database. In one embodiment the invention is directed to a method for determining an amino acid sequence of a peptide which comprises: (a) obtaining a suitable fragmentation mass spectrum having a plurality of peaks for the peptide; (b) removing the peaks due to C13 isotopes from the spectrum and applying an appropriate intensity threshold to the remaining peaks; (c) selecting a suitable peak as a starting point and determining mass differences in the spectrum that corresponds to an amino acid residue mass difference; (d) sequentially determining each subsequent amino acid residue mass difference from the starting peak; (e) repeating the process of steps (c) and (d) for additional peaks so as to obtain a non-redundant set of proposed sequences; and (f) ranking the proposed sequences and preparing a consensus sequence corresponding to the amino acid sequence of the peptide.

In this invention the peptide may be obtained by selective cleavage of a polypeptide. If so, then any proposed amino acid sequences inconsistent with the selective cleavage can be removed to obtain a revised set of proposed sequences prior to ranking the proposed sequences. In the method the selective cleavage may be enzymatic cleavage. Preferably, the selective enzymatic cleavage is cleavage with arginine endopeptidase (ArgC); asparatic acid endopeptidase N (aspN); chymotrypsin; glutamic

acid endopeptidase C (gluC); lysine endopeptidase C (lysC); trypsin; or V8 endopeptidase. More preferable enzymatic cleavage is cleavage with trypsin.

In one embodiment the suitable peak as the starting point is a high mass peak. In addition, the sequential determination of step (d) may be repeated until nearly every possible amino acid residue mass difference is investigated. Preferably, the sequential determination of step (d) is repeated until every possible amino acid residue mass difference is investigated.

In one embodiment of the invention the amino acid sequence is a partial amino acid sequence. Preferably the partial amino acid sequence is a trimer. Alternatively, the amino acid sequence is a complete amino acid sequence.

Typically, the suitable fragmentation mass spectrum is obtained from a suitable mass spectrometer. Processes which produce fragmentation that can be used in generating a suitable fragmentation mass spectrum, include but are not limited to, collisionally-induced dissociation, collisionally-activated dissociation, post-source decay, surface-induced dissociation, and in-source fragmentation. A suitable mass spectrum can be generated from tandem mass spectrometry or multiple stages of mass spectrometry. In one embodiment, the suitable mass spectrum is obtained from a tandem mass spectrometer, preferably, a quadrupole tandem time of flight mass spectrometer (Q-TOF). In principle, other instrument types and configurations can be used, provided there is sufficient number of the required suite of sequencing ions to generate sequence information. These include, but not limited to, tandem magnetic sector instruments, Fourier-transform ion cyclotron resonance mass spectrometer, triple quadrupole ion trap mass spectrometers, tandem time-of-flight mass spectrometers (TOF-TOF). The configuration of laser-desorption matrix-assisted ionization-triple quadrupole mass analyser-collision cell-TOF can also be used to obtain a peptide fragment spectrum. Ionisation processes which can be used, include but not limited to electrospray ionisation, nanoflow electrospray ionisation, matrix-assisted laser desorption ionisation, plasma desorption ionisation, fast atom bombardment, and field desorption.

One aspect of the preferred embodiment of the HOPS method is the ability to determine whether the difference in mass-to-charge ratio (*m/z*) between two product ions in the MS/MS spectrum of a peptide is consistent with the mass of a known amino acid group. In one embodiment of the invention the product ions used for the sequential determination of the amino acid sequence is the y-ion series, alternatively a b-ion series

may be used or a combination of both. In the method preferably about five to about fifty product ions are selected as suitable peaks as starting points for the sequential determination of the amino acid mass residue difference, preferably about ten to about thirty peaks are selected as suitable peaks as starting points for the sequential determination of the amino acid mass residue difference. The masses of the amino acid groups are assumed to be monoisotopic. As amino acid masses can be specified up to 5 decimal places in units of Da, any experimental uncertainly in their mass is entirely negligible compared to the experimental uncertainty in the experimentally obtained *m/z* values. As the *m/z* values of product ion peaks are subject to experimental error derived from the mass resolution of the mass spectrometer, minimum and maximum values need to be computed for the *m/z* difference. If the mass of the amino acid falls within this range, then there is deemed to be a correspondence between the experimentally observed peaks and the mass of a particular amino acid residue.

To define an embodiment of this process we first consider the physical model of the data representing the fragmentation spectrum. The fragmentation spectrum is represented from an external data file containing the following information:

1. The mass of the peptide ion analysed (termed the precursor ion) to produce the MS/MS spectrum, and the charge state of the precursor ion.
2. A list of paired values of mass-to-charge ratio (*m/z*) and intensity, corresponding to the peaks in spectrum. These are arranged in order of ascending *m/z* values.

As the mass spectrometry instrumentation used to produce the spectrum has a finite mass resolution, the *m/z* co-ordinate is subject to experimental uncertainly and is therefore not a precise value. In the particular implementation of the model, the experimental uncertainly is assumed to result in a normal statistical distribution. The given *m/z* value from the peak table is then taken to be the mean value, and the experimental mass resolution of the instrument is assumed to be equal to the standard deviation from the mean. The experimental uncertainly is then calculated as being ±2 standard deviations from the mean value, *i.e.* the 95% confidence limit of the normal distribution.

Thus, in order to calculate whether the difference between two particular *m/z* values corresponds to the mass of an amino acid group, a mass range that incorporates the uncertainties in both *m/z* values is calculated which is dependent on the mass

resolution of the type of instrument used. In a preferred embodiment the minimum instrument resolution required by the HOPS sequencing method for unambiguous interpretation of peptide fragmentation spectra may be determined by the following calculation. For the mass spectrometer used (in this case, the Q-tof), each peak has a finite width and this width acts as the error for an individual peak. This error is compounded for the difference in m/z values of two peaks. For two peaks we calculate the compound error E from the components E_1 and E_2 due to the two peaks:

$$\Delta E = \sqrt{\left(E_1^2 + E_2^2\right)} \qquad (1)$$

To a first approximation, we assume that the width of each peak is the same:

$$\Delta E = \sqrt{2E^2} \qquad (2)$$

To obtain an unambiguous determination of the correct amino acid residue, the compound error on the difference between the two peaks must be less than 1 Dalton. Therefore for a correct HOPS read:

$$\sqrt{2E^2} < 1 \qquad (3)$$

We can write the error, E, at mass, M, in terms of the mass resolution of the mass spectrometer, R:

$$E = \frac{M}{R} \qquad (4)$$

and so condition (3) becomes

$$M < \frac{R}{\sqrt{2}} \qquad (5)$$

Therefore, in this embodiment, as long as condition (5) is met for any mass, M, within a mass spectrometer of resolution, R, then HOPS will be able to unambiguously interpret the spectrum. In a range 0-2000 Da, a minimum resolution of ~2800 (full width half maximum height) at M=2000 Da is preferable to be able to carry out a HOPS

analysis at 2000 Da. In another embodiment, the suitable fragmentation mass spectrum should have a minimum resolution of at least 5600 full width half peak height for peptides with a molecular weight up to about 4000 daltons.

In the determination of whether the difference in mass between two peaks in the spectrum correspond to the mass of an amino acid residue, the set of amino acid residues considered is not limited to the twenty naturally-occurring amino acid residues but may include modified amino acids. The method may be applied to the sequential determination of peptides having modified amino acids where these modified amino acids are incorporated in the set of amino acid residues considered in determining whether the mass difference between two peaks corresponds to the mass of an amino acid residue. In one embodiment of the invention, HOPS can be used to determine sequence information for peptides comprising at least one post-translationally modified amino acid. Alternatively, HOPS can be used to determine sequence information for peptides comprising a plurality of post-translationally modified amino acids. In one embodiment of the invention the polypeptide is a polypeptide that has not been disclosed in a publically available database.

In one embodiment, an amino acid sequence determined by HOPS may be compared to conceptual polypeptides translated from nucleotide sequences in a database and the results of the comparison are used to identify expressed regions in a nucleotide sequence. In a further embodiment, results of the comparison may be used to identify an error due to a nucleotide mutation or due to nucleotide sequencing. These methods may be repeated for a plurality of amino acid sequences determined by HOPS in order to identify more than one expressed region in a nucleotide, more than one nucleotide mutation or more than one error due to nucleotide sequencing. The database of conceptual polypeptides translated nucleotide sequences may be derived from genomic sequences, more preferably from the human genome.

In one embodiment, a partial amino acid sequence of a peptide determined by HOPS is combined with the total mass of the peptide to obtain a first mass and a second mass for the peptide and the first mass, partial sequence, second mass combination is used to identify the polypeptide in a database. The composition of the database may include, but is not limited to, proteins, polypeptides, peptides, or conceptual polypeptides translated from nucleotide sequences, or any combination thereof. Preferably, neither the first mass nor the second mass is a single amino acid residue mass, and the partial amino

acid sequence used is a trimer. Preferably, the total mass of the peptide is determined to an error of measurement of about 200 parts per million (ppm) or less. Methods of mass analysis suitable for determining the total mass of a peptide include, but are not limited to: time-of-flight, Fourier transform ion cyclotron resonance, quadrupole, ion trap, and magnetic sector analysis. A preferred mass spectrometer for determination of the total mass of a peptide is a matrix assisted laser desorption time-of-flight (MALDI-TOF) mass spectrometer. In a further embodiment, the first mass, partial amino acid sequence, second mass combination compared with conceptual polypeptides from nucleotide sequences to identify expressed regions within nucleotide sequences. In yet a further embodiment, results of the comparison may be used to identify an error due to a nucleotide mutation or due to nucleotide sequencing. These methods may be repeated for a plurality of first mass, partial amino acid sequence, second mass combinations determined by HOPS in order to identify more than one expressed region in a nucleotide, more than one nucleotide mutation or more than one error due to nucleotide sequencing.

The invention also provides a method of identifying a polypeptide(s) in a database based on a peptide obtained by selective cleavage of an unknown polypeptide which comprises: (a) obtaining a suitable first mass, a suitable sequence of three or more amino acids, and a second mass from the peptide; (b) prepare a permuted search set based on the suitable amino acid sequence by including possible sequences due isobaric amino acids and amino acids ambiguous due to the resolution of the instrument; (c) comparing the first mass, the permuted search set and the second mass to the database so as to obtain a set of possible polypeptides corresponding to the unknown polypeptide; and (d) removing from the set of possible polypeptides any polypeptide inconsistent with the selective cleavage method so as to uniquely identify the polypeptide(s). Preferably, the suitable sequence of three or more amino acids is chosen to exclude specific amino acid combinations that appear with a high frequency in the database.

In one embodiment of the invention the database is a database of protein sequences. Alternatively, the database is a database of peptide sequences, a database of peptide sequences prepared from a database of protein sequences, or a database of peptide sequences prepared from a database of nucleic acid sequences. Preferably, the peptide database is constrained to remove any sequences that include a residue that appears adjacent to a stop codon in the nucleotide database. The nucleic acid database

may contains over 200,000 sequences, over 500,000 sequences, over 1,000,000 sequences, over 10,000,000 sequences or over 100,000,000 sequences.

The permuted search set may be selected from one or more of the following permutations: arginine permuted with (valine and glycine) and vis versa; asparagine permuted with two glycines and vis versa; leucine permuted with isoleucine and vis versa; lysine permuted with glutamine and vis versa; phenylalanine permuted with oxidized methionine and vis versa; tryptophan permuted with (alanine and aspartic acid) and vis versa; tryptophan permuted with (glycine and glutamic acid) and vis versa; or tryptophan permuted with (valine and serine) and vis versa.

The suitable amino acid sequence may be obtained by the HOPS method described above. In one embodiment of the invention the database is a nucleotide database and the polypeptide has a predicted polypeptide sequence encoded by a nucleotide sequence; and mapping the sequence of the peptide onto the nucleotide sequence; and then mapping either the consensus or the proposed sequence for at least one additional peptide from the polypeptide on to the nucleotide sequence so as to identify portions the nucleotide sequence that are expressed. Preferably a plurality of additional peptides are mapped on to the nucleotide sequence. In addition a total mass for at least one additional peptide may be mapped on to the nucleotide sequence, preferably a plurality of total masses for a plurality of additional peptides are mapped on to the nucleotide sequence. The methods herein may be used for the identification of an intron or exon boundary, expressed sequences or to identify protein isoforms or post translational modifications within proteins.

In an alternative embodiment of the invention it provides a method for identifying expressed regions in a nucleotide sequence present in a database which comprises:(a) obtaining a peptide by selective cleavage of a polypeptide; (b) obtaining a suitable first mass, a suitable sequence of three or more amino acids, and a second mass for the peptide; (c) preparing a permuted search set based on the suitable amino acid sequence by including possible sequences due isobaric amino acids and amino acids ambiguous due to the resolution of the instrument; (d) comparing the first mass, the permuted search set and the second mass to the database so as to obtain a set of possible polypeptides corresponding to the polypeptide; (e) removing from the set of possible polypeptides any polypeptide inconsistent with the selective cleavage method so as to identify the nucleotide sequence encoding the polypeptide in the database and the nucleotide

sequence of the peptide encoded therein; (f) repeating steps (a) through (e) for at least one more peptide obtained from the polypeptide; and (g) analysing the results of steps (e) and (f) so as to identify expressed in a nucleotide sequence. In a further embodiment, other expressed regions within the identified nucleotide are determined by mapping the total masses of other peptides onto the nucleotide. Preferably the total masses for the peptides are determined within an error of mass measurement of 1ppm or less. In another embodiment these methods can be used to identify post-translational modifications present in the protein encoded by the nucleotide. In a further embodiment these methods can be used to determine single nucleotide polymorphisms, or to identify errors due to nucleotide mutations or due to nucleotide sequencing.

In one embodiment of the invention the amino acid sequence of the peptide is used to diagnose a disease. Here, the peptide may be used to diagnose a disease associated with a specific post-translational modification. Examples of such diseases include, but are not limited to, diseases associated with aging and glycosylation, prion associated diseases, metabolic disorders.

Examples of post-translational modifications that may be identified, sequenced or mapped using the methods described herein, include but are not limited to: N-formyl-L-methionine ; L-selenocysteine; L-cystine; L-erythro-beta-hydroxyasparagine; L-erythro-beta-hydroxyaspartic acid; 5-hydroxy-L-lysine; 3-hydroxy-L-proline; 4-hydroxy-L-proline; 2-pyrrolidone-5-carboxylic acid; L-gamma-carboxyglutamic acid; L-aspartic 4-phosphoric anhydride; S-phospho-L-cysteine; 1'-phospho-L-histidine; 3'-phospho-L-histidine; O-phospho-L-serine; O-phospho-L-threonine; O4'-phospho-L-tyrosine; 2'-[3-carboxamido-3-(trimethylammonio)propyl]-L-histidine; N-acetyl-L-alanine; N-acetyl-L-aspartic acid; N-acetyl-L-cysteine; N-acetyl-L-glutamic acid; N-acetyl-L-glutamine; N-acetylglycine; N-acetyl-L-isoleucine; N2-acetyl-L-lysine; N-acetyl-L-methionine; N-acetyl-L-proline; N-acetyl-L-serine; N-acetyl-L-threonine; N-acetyl-L-tyrosine; N-acetyl-L-valine; N6-acetyl-L-lysine; S-acetyl-L-cysteine; N-formylglycine; D-glucuronyl-N-glycine; N-myristoyl-glycine; N-palmitoyl-L-cysteine; N-methyl-L-alanine; N,N,N-trimethyl-L-alanine; N-methylglycine; N-methyl-L-methionine; N-methyl-L-phenylalanine; N,N-dimethyl-L-proline; omega-N,omega-N'-dimethyl-L-arginine; omega-N,omega-N-dimethyl-L-arginine; omega-N-methyl-L-arginine; N4-methyl-L-asparagine; N5-methyl-L-glutamine; L-glutamic acid 5-methyl ester; 3'-methyl-L-histidine; N6,N6,N6-trimethyl-L-lysine; N6,N6-dimethyl-L-lysine; N6-

methyl-L-lysine; N6-palmitoyl-L-lysine; N6-myristoyl-L-lysine; O-palmitoyl-L-threonine; O-palmitoyl-L-serine; L-alanine amide; L-arginine amide; L-asparagine amide; L-aspartic acid 1-amide; L-cysteine amide; L-glutamine amide; L-glutamic acid 1-amide; glycine amide; L-histidine amide; L-isoleucine amide; L-leucine amide; L-lysine amide; L-methionine amide; L-phenylalanine amide; L-proline amide; L-serine amide; L-threonine amide; L-tryptophan amide; L-tyrosine amide; L-valine amide; L-cysteine methyl disulfide; S-farnesyl-L-cysteine; S-12-hydroxyfarnesyl-L-cysteine; S-geranylgeranyl-L-cysteine; L-cysteine methyl ester; S-palmitoyl-L-cysteine; S-diacylglycerol-L-cysteine; S-(L-isoglutamyl)-L-cysteine; 2'-(S-L-cysteinyl)-L-histidine; L-lanthionine; meso-lanthionine; 3-methyl-L-lanthionine; 3'-(S-L-cysteinyl)-L-tyrosine; N6-carboxy-L-lysine; N6-1-carboxyethyl-L-lysine; N6-(4-amino-2-hydroxybutyl)-L-lysine; N6-biotinyl-L-lysine; N6-lipoyl-L-lysine; N6-pyridoxal phosphate-L-lysine; N6-retinal-L-lysine; L-allysine; L-lysinoalanine; N6-(L-isoglutamyl)-L-lysine; N6-glycyl-L-lysine; N-(L-isoaspartyl)-glycine; pyruvic acid; L-3-phenyllactic acid; 2-oxobutanoic acid; N2-succinyl-L-tryptophan; S-phycocyanobilin-L-cysteine; S-phycoerythrobilin-L-cysteine; S-phytochromobilin-L-cysteine; heme-bis-L-cysteine; heme-L-cysteine; tetrakis-L-cysteinyl iron; tetrakis-L-cysteinyl diiron disulfide; tris-L-cysteinyl triiron trisulfide; tris-L-cysteinyl triiron tetrasulfide; tetrakis-L-cysteinyl tetrairon tetrasulfide; L-cysteinyl homocitryl molybdenum-heptairon-nonasulfide; L-cysteinyl molybdopterin; S-(8alpha-FAD)-L-cysteine; 3'-(8alpha-FAD)-L-histidine; O4'-(8alpha-FAD)-L-tyrosine; L-3',4'-dihydroxyphenylalanine; L-2',4',5'-topaquinone; L-tryptophyl quinone; 4'-(L-tryptophan)-L-tryptophyl quinone; O-phosphopantetheine-L-serine; N4-glycosyl-L-asparagine; S-glycosyl-L-cysteine; O5-glycosyl-L-hydroxylysine; O-glycosyl-L-serine; O-glycosyl-L-threonine; 1'-glycosyl-L-tryptophan; O4'-glycosyl-L-tyrosine; N-asparaginyl-glycosylphosphatidylinositolethanolamine;
N-aspartyl-glycosylphosphatidylinositolethanolamine;
N-cysteinyl-glycosylphosphatidylinositolethanolamine;
N-glycyl-glycosylphosphatidylinositolethanolamine;
N-seryl-glycosylphosphatidylinositolethanolamine;
N-alanyl-glycosylphosphatidylinositolethanolamine;
N-seryl-glycosylsphingolipidinositolethanolamine; O-(phosphoribosyl dephospho-coenzyme A)-L-serine; omega-N-(ADP-ribosyl)-L-arginine; S-(ADP-ribosyl)-L-cysteine; L-glutamyl 5-glycerylphosphorylethanolamine; S-sulfo-L-cysteine; O4'-sulfo-

L-tyrosine; L-bromohistidine; L-2'-bromophenylalanine; L-3'-bromophenylalanine; L-4'-bromophenylalanine; 3',3",5'-triiodo-L-thyronine; L-thyroxine; L-6'-bromotryptophan; dehydroalanine; (Z)-dehydrobutyrine; dehydrotyrosine; L-seryl-5-imidazolinone glycine; L-3-oxoalanine; lactic acid; L-alanyl-5-imidazolinone glycine; L-cysteinyl-5-imidazolinone glycine; D-alanine; D-allo-isoleucine; D-methionine; D-phenylalanine; D-serine; D-asparagine; D-leucine; D-tryptophan; L-isoglutamyl-polyglycine; L-isoglutamyl-polyglutamic acid; O4'-(phospho-5'-adenosine)-L-tyrosine; S-(2-aminovinyl)-D-cysteine; L-cysteine sulfenic acid; S-glycyl-L-cysteine; S-4-hydroxycinnamyl-L-cysteine; chondroitin sulfate D-glucuronyl-D-galactosyl-D-galactosyl-D-xylosyl-L-serine; dermatan 4-sulfate D-glucuronyl-D-galactosyl-D-galactosyl-D-xylosyl-L-serine; heparan sulfate D-glucuronyl-D-galactosyl-D-galactosyl-D-xylosyl-L-serine; N6-formyl-L-lysine; O4-glycosyl-L-hydroxyproline; O-(phospho-5'-RNA)-L-serine; L-citrulline; 4-hydroxy-L-arginine; N-(L-isoaspartyl)-L-cysteine; 2'-alpha-mannosyl-L-tryptophan; N6-mureinyl-L-lysine; 1-chondroitin sulfate-L-aspartic acid ester; S-(6-FMN)-L-cysteine; 1'-(8alpha-FAD)-L-histidine; omega-N-phospho-L-arginine; S-diphytanylglycerol diether-L-cysteine; alpha-1-microglobulin-Ig alpha complex chromophore; bis-L-cysteinyl bis-L-histidino diiron disulfide; hexakis-L-cysteinyl hexairon hexasulfide; N6-(phospho-5'-adenosine)-L-lysine; N6-(phospho-5'-guanosine)-L-lysine; L-cysteine glutathione disulfide; S-nitrosyl-L-cysteine; N4-(ADP-ribosyl)-L-asparagine; L-beta-methylthioaspartic acid; 5'-(N6-L-lysine)-L-topaquinone; S-methyl-L-cysteine; 4-hydroxy-L-lysine; N4-hydroxymethyl-L-asparagine; O-(ADP-ribosyl)-L-serine; L-cysteine oxazolecarboxylic acid; L-cysteine oxazolinecarboxylic acid; glycine oxazolecarboxylic acid; glycine thiazolecarboxylic acid; L-serine thiazolecarboxylic acid; L-phenyalanine thiazolecarboxylic acid; L-cysteine thiazolecarboxylic acid; L-lysine thiazolecarboxylic acid; O-(phospho-5'-DNA)-L-serine; keratan sulfate D-glucuronyl-D-galactosyl-D-galactosyl-D-xylosyl-L-threonine; L-selenocysteinyl molybdopterin guanine dinucleotide; O4'-(phospho-5'-RNA)-L-tyrosine; 3-(3'-L-histidyl)-L-tyrosine; L-methionine sulfone; dipyrrolylmethanemethyl-L-cysteine; S-(2-aminovinyl)-3-methyl-D-cysteine; O4'-(phospho-5'-DNA)-L-tyrosine; O-(phospho-5'-DNA)-L-threonine; O4'-(phospho-5'-uridine)-L-tyrosine; N-(L-glutamyl)-L-tyrosine; S-phycobiliviolin-L-cysteine; phycoerythrobilin-bis-L-cysteine; phycourobilin-bis-L-cysteine; N-L-glutamyl-poly-L-glutamic acid; L-cysteine sulfinic acid; L-3',4',5'-trihydroxyphenylalanine; O-(sn-1-glycerophosphoryl)-L-serine; 1-

thioglycine; heme P460-bis-L-cysteine-L-tyrosine; O-(phospho-5'-adenosine)-L-threonine; tris-L-cysteinyl-L-cysteine persulfido-bis-L-glutamato-L-histidino tetrairon disulfide trioxide; L-cysteine persulfide; 3'-(1'-L-histidyl)-L-tyrosine; heme P460-bis-L-cysteine-L-lysine; 5-methyl-L-arginine; 2-methyl-L-glutamine; N-pyruvic acid 2-iminyl-L-cysteine; N-pyruvic acid 2-iminyl-L-valine; heme-L-histidine; S-selenyl-L-cysteine; N6-methyl-N6-poly(N-methyl-propylamine)-L-lysine; hemediol-L-aspartyl ester-L-glutamyl ester; hemediol-L-aspartyl ester-L-glutamyl ester-L-methionine sulfonium; L-cysteinyl molybdopterin guanine dinucleotide; trans-2,3-cis-3,4-dihydroxy-L-proline; pyrroloquinoline quinone; tris-L-cysteinyl-L-N1'-histidino tetrairon tetrasulfide; tris-L-cysteinyl-L-N3'-histidino tetrairon tetrasulfide; tris-L-cysteinyl-L-aspartato tetrairon tetrasulfide; N6-pyruvic acid 2-iminyl-L-lysine; tris-L-cysteinyl-L-serinyl tetrairon tetrasulfide; bis-L-cysteinyl-L-N3'-histidino-L-serinyl tetrairon tetrasulfide; O-octanoyl-L-serine. One of ordinary skill in the art would readily recognize that other post-translational modifications occur. One skilled in the art would readily recognise the ability to use the molecular weight for these modifications and incorporate them into the method described and claimed herein.

One of ordinary skill will readily recognize that the methods described herein may be used to detect a variety of post-translational modifications relevant to basic research or to the clinical diagnosis of disease. Examples of the types of PTMs that may be analyzed using the methods described herein include, but are not limited to alkylation, see e.g. Saragoni et al. (2000) Differential association of tau with subsets of microtubules containing posttranslationally-modified tubulin variants in neuroblastoma cells. Neurochem. Res. 25:59-70; Fanapour et al. (1999) Hyperhomocysteinemia: an additional cardiovascular risk factor. WMJ 98:51-4; Raju et al. (1997) N-Myristoyltransferase overexpression in human colorectal adenocarcinomas. Exp. Cell Res. 235:145-54; Zhao et al. (2000) Palmitoylation of apolipoprotein B is required for proper intracellular sorting and transport of cholesteroyl esters and triglycerides. Mol. Biol. Cell 11:721-34; or Seabra MC (1996) Nucleotide dependence of Rab geranylgeranylation. Rab escort protein interacts preferentially with GDP-bound Rab. J. Biol. Chem. 271:14398-404, the contents of which are hereby incorporated in their entirety.

Examples of phosphorylation include, but are not limited to, Vanmechelen et al. (2000) Quantification of tau phosphorylated at threonine 181 in human cerebrospinal fluid: a sandwich ELISA with a synthetic phosphopeptide for standardization. Neurosci.

Lett. 285:49-52; Lutz et al. (1994) Characterization of protein serine/threonine phosphatases in rat pancreas and development of an endogenous substrate-specific phosphatase assay. Pancreas 9:418-24; Gitlits et al. (2000) Novel human autoantibodies to phosphoepitopes on mitotic chromosomal autoantigens (MCAs). J. Investig. Med. 48:172-82; or Quin and McGuckin (2000) Phosphorylation of the cytoplasmic domain of the MUC1 mucin correlates with changes in cell-cell adhesion. Int. J. Cancer 87:499-506, the contents of which are hereby incorporated in their entirety.

A example of sulfation includes, but is not limited to, Manzella et al. (1995) Evolutionary conservation of the sulfated oligosaccharides on vertebrate glycoprotein hormones that control circulatory half-life. J. Biol. Chem. 270S:21665-71, the contents of which is hereby incorporated in its entirety.

Examples of oxidation/reduction include, but are not limited to, Magsino et al. (2000) Effect of triiodothyronine on reactive oxygen species generation by leukocytes, indices of oxidative damage, and antioxidant reserve. Metabolism 49:799-803; or Stief et al. (2000) Singlet oxygen inactivates fibrinogen, factor V, factor VIII, factor X, and platelet aggregation of human blood. Thromb. Res. 97:473-80, the contents of which are hereby incorporated in their entirety.

Examples of ADP-ribosylation include, but are not limited to, Galluzzo et al. (1995) Involvement of CD44 variant isoforms in hyaluronate adhesion by human activated T cells. Eur. J. Immunol. 25:2932-9; or Thraves et al. (1986) Differential radiosensitization of human tumour cells by 3-aminobenzamide and benzamide: inhibitors of poly(ADP-ribosylation). Int. J. Radiat. Biol. Relat. Stud. Phys. Chem. Med. 50:961-72, the contents of which are hereby incorporated in their entirety.

A example of hydroxylation includes, but is not limited to, Brinckmann et al. (1999) Overhydroxylation of lysyl residues is the initial step for altered collagen cross-links and fibril architecture in fibrotic skin. J. Invest. Dermatol. 113:617-21, the contents of which is hereby incorporated in its entirety.

Examples of glycosylation include, but are not limited to, Johnson et al. (1999) Glycan composition of serum alpha-fetoprotein in patients with hepatocellular carcinoma and non-seminomatous germ cell tumour. Br. J. Cancer 81:1188-95; Fulop et al. (1996) Species-specific alternative splicing of the epidermal growth factor-like domain 1 of cartilage aggrecan. Biochem. J. 319:935-40; Dow et al. (1994) Molecular correlates of spinal cord repair in the embryonic chick: heparan sulfate and chondroitin sulfate

proteoglycans. Exp. Neurol. 28:233-8; Kelly et at. (1993) RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc. J. Biol. Chem. 268:10416-24; Goss et al. (1995) Inhibitors of carbohydrate processing: A new class of anticancer agents. Clin. Cancer Res. 1:935-44; or Sleat et al. (1998) Specific alterations in levels of mannose 6-phosphorylated glycoproteins in different neuronal ceroid lipofuscinoses. Biochem. J. 334:547-51, the contents of which are hereby incorporated in their entirety.

An example of glucosylphosphatidylinositide addition includes, but is not limited to, Poncet et al. (1996) CD24, a glycosylphosphatidylinositol-anchored molecules is transiently expressed during the development of human central nervous system and is a marker of human neural cell lineage tumors. Acta Neuropathol. (Berl.) 91:400-8, the contents of which is hereby incorporated in its entirety.

An example of ubiquitination includes, but is not limited to,Chu et al. (2000) Ubiquitin immunochemistry as a diagnostic aid for community pathologists evaluating patients who have dementia. Mod. Pathol. 13:420-6, the contents of which is hereby incorporated in its entirety.

An example of translocation includes, but is not limited to, Reddy et al. (1999) Recent advances in understanding the pathogenesis of Huntington's disease. Trends Neurosci. 22:248-55, the contents of which is hereby incorporated in its entirety.

An example of detection of an artificial modification (e.g., biotinylation, cross-linking, photoaffinity labeling) includes, but is not limited to, Romero et al. (1993) Differential T cell receptor photoaffinity labeling among H-2Kd restricted cytotoxic T lymphocyte clones specific for a photoreactive peptide derivative. Labeling of the alpha-chain correlates with J alpha segment usage. J. Exp. Med. 177:1247-56, the contents of which is hereby incorporated in its entirety.

The methods described herein may also be used to detect proteolytic processing or changes associated with transcription or genetic changes. Examples of proteolytic processing include, but are not limited to, Kurahara et al. (1999) Expression of MMPS, MT-MMP, and TIMPs in squamous cell carcinoma of the oral cavity: correlations with tumor invasion and metastasis. Head Neck 21:627-38; or Thorgeirsson et al. (1994) Tumor invasion, proteolysis, and angiogenesis. J. Neurooncol. 18:89-103, the contents of which are hereby incorporated in their entirety.

Examples of primary sequence variability (e.g., mRNA splicing variability, gene mutation) include, but are not limited to, Fulop et al. (1996) Species-specific alternative splicing of the epidermal growth factor-like domain 1 of cartilage aggrecan. Biochem. J. 319:935-40; or Bergquist et al (2000) Rapid Method to Characterize Mutations in Transthyretin in Cerebrospinal Fluid from Familial Amyloidotic Polyneuropathy Patients by Use of Matrix-assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. Clin. Chem. 46:1293-1300, the contents of which are hereby incorporated in their entirety.

FIG. 2 shows an overview of a preferred embodiment of the invention for the identification and editing of expressed genomic sequences and the characterisation of their post-translational modifications. The protein to be identified is first digested with a specific endoprotease such as trypsin, to cleave the protein into constitutive peptides. Although one embodiment of the present invention describes the use of tryptic digestion, other enzymes with sufficiently restrictive cleavage patterns may also be used, including but not limited to Lys C, AspN, Glu-C. In the method of the present invention, digestion of the protein is carried our under conditions to maximise the number of peptides which contain only C-terminal Arginine and Lysine residues. In the preferred embodiment, protein features are excised from either a one or two-dimensional gel by a software-driven robotic cutter as described in U.S. patent number 6,064,754, the contents of which is incorporated in its entirety. The gel pieces are then subjected to in situ proteolysis in an OGS ChemStation robot and using a modification of the following manual method (Page et al. (1999) *Proc. Natl. Acad. Sci.* 96: 12589-12594).

In a preferred embodiment the robotically cut gel plugs are washed with 50µl of 100mM ammonium bicarbonate to each sample. After standing for 10 minutes at ambient temperature, the liquid is removed. Acetonitrile (50µl) is added to each tube. Let stand for 10 minutes at ambient temperature and with manual agitation for 5 minutes. The samples are dried by centrifugal evaporation for 10 minutes with no heating. Fifty µL of ammonium bicarbonate (100 mM) is added to each tube. Let stand at ambient temperature for 10 minutes and remove the liquid. Acetonitrile (50µl) is added to each tube. Let stand for 10 minutes at ambient temperature, remove the liquid, and dry the sample by centrifugal evaporation for 10 minutes with no heating. Porcine trypsin

(133ng in 5 uL) is added to each sample. After 5 minutes at room temperature, 5µl of the 66.5 ng trypsin is added to each sample. The samples are incubated at 40°C for 2 hr in an oven and after cooling at room temperature for 5 minutes are centrifuged for 1 minute at 13,000 rpm. All the liquid (peptide pool) surrounding the gel piece is removed and dispensed into clean 0.5ml test tubes for mass spectrometric analysis.

Trypsin cleaves specifically at the carboxyl side of lysine (Lys) and arginine (Arg) residues, so that the tryptic digest peptide fragments generated should have a Lys or Arg as the C-terminal amino acid, unless the peptide fragment was obtained from the C-terminal of the protein. Similarly, the amino acid directly preceding the N-terminal amino acid of the peptide fragment in the protein should also be a Lys or Arg, unless the peptide was obtained from the N-terminal of the protein.

The mixture of peptides from individual proteins can be analysed by mass spectrometry without any prior separation (as shown in **FIG.** 2, step 2) or can be separated into individual peptides using well known chromatographic methods. In a preferred embodiment of this invention, the tryptic peptides are initially analysed using matrix-assisted laser-desorption time-of-flight mass spectrometry with delayed extraction and a reflectron in the time-of-flight chamber (MALDI-TOF). This instrument configuration is used to determine accurately the molecular weights (< 100 parts-per-million (ppm)) of the modified and unmodified peptides and to increase the throughput of subsequent mass spectrometric analyses. Other methods of mass analysis capable of mass measurement within an error of 100ppm or less include but are not limited to: time-of-flight, Fourier transform ion cyclotron resonance, quadrupole, ion trap, and magnetic sector analysis.

In order to determine the amino acid sequence of the tryptic peptides, the peptides are then analysed by tandem mass spectrometry to obtain a fragmentation or MS/MS spectrum of the peptide (**FIG.** 2, step 3). In a preferred embodiment a resolution of greater than 4000 (peak width at half maximum height) and a mass accuracy of less than 50 ppm (parts-per-million) is preferred. In the method of the present invention, a hybrid tandem mass spectrometer with a triple quadrupole mass analyser as M1 and a time-of-flight mass analyser as M2, is used for tandem mass spectrometry analysis. In a preferred embodiment of the present method, tandem mass spectrometry analysis is carried out on doubly charged peptide ions, although the method is not limited to the analysis of peptide ions of other charge states. The quadrupole mass analyser is set to

allow transmission of ions with an *m/z* equal to the half the mass of the doubly-charged quasi molecular ion of a putative peptide signal observed in the MALDI-TOF analysis (termed the 'precursor' or 'daughter' ion). The peptide ion beam passes into the collision cell where the peptide ions are subjected to fragmentation via collision-induced dissociation. This is achieved through the application of a voltage on the collision cell and by the introduction of an inert gas. The fragment ions produced (termed the 'product' or 'daughter' ions) in addition to any remaining intact precursor ions are then pulsed into the second mass analyser (TOF) where their masses are determined by their arrival time at the detector. The resulting fragment or MS/MS spectrum is typically represented by a two-dimensional graph with intensity on the y-axis, and mass-to-charge ratio (*m/z*) on the x-axis. In addition, the peak height of each defined *m/z* value is measure as an ion intensity. The output of the sequencing process of HOPS produces an array of sequences, termed the 'consensus' sequence (**FIG.** 2., step 4). All sequences read by HOPS are retained and mapped onto identified protein sequences or translated nucleotide sequences at a later step (**FIG.** 2., step 7).

An overview of one embodiment of the HOPS process which produces the consensus sequence from which the search string is constructed shown in **FIG.** 3. The search string is constructed so as to produce three distinct pieces of information unique to the spectrum being analysed, namely a partial amino acid sequence, an N-terminal mass (denoted M1) and a C-terminal mass (denoted M2). The N-terminal mass or M1 is the mass between the start of the partial amino acid sequence and the N-terminus of the candidate peptide, and the C-terminal mass or M2 is the mass between the end of the partial amino acid sequence and the C-terminus of the candidate peptide.

In a preferred embodiment, the HOPS algorithm incorporates the following components:

1 A 'Peak table' object. This object incorporates an expandable array of *m/z* and intensity paired values sorted in order of increasing *m/z*. Also incorporated are methods for the extraction of a single co-ordinate (*m/z*, intensity) of paired values at any point in the array, and methods for the complete of particular paired values should that particular entry in the peak table object be determined to be invalid

2 A 'Walk' object. This object contains several expandable arrays:

(a) A *m/z* value array. This corresponds to peaks from the peak table object which can be assigned to particular amino acid masses;

(b) A 'b-ion' Boolean flag array. This is set true if the peak in the array in (a) above has a complementary b-ion identified in the peak table;

(c) A 'a-ion' Boolean flag array. This is set true if the peak in (a) above has a signal with an *m/z* decrement of 27.997Da from a b ion signal (corresponding to the loss of a carbon and oxygen from a b ion);

In addition, the walk array contains integer values to store the scoring of the values in the arrays compared with the values in the peak table.

3 The 'Stack'. This is an expandable array of walk objects. Each walk object is identified by an index number which is its relative position on the stack.

4 An Amino acid object. This contains a list of the masses of the amino acid masses applicable in any study. This may be confined to the 20 naturally occurring amino acids (listed in table 1), or may include masses corresponding to modifications of these amino acids caused post-translational modifications. The object also contains an identifying symbol for each of the amino acids.

In addition, the algorithm maintains a variable pointing to the index number of the walk currently under consideration.

In a preferred embodiment, all sequence reads from the spectral input are kept as possibilities without any pruning or rejection. All possibilities are later reviewed and ranked, and the output sequence is deduced through a consensus process. In a preferred aspect of the present invention, the HOPS method is implemented to obtain highly specific sequence information to be used to search databases comprising proteins, polypeptides, peptides or conceptual polypeptides translated from nucleotide sequences, or any combination thereof. The HOPS method, however is not limited to use with database searching and can also be used as a method for the interpretation of fragmentation mass spectra of peptides without any application of the resulting sequence information to database searching. The method can also be used for the sequence determination of peptides obtained from means other than tryptic digestion and can also be used for the sequence determination of peptides containing post-translational modifications.

FIG. 4 is a flow diagram showing the process by which the edited Peak Table is calculated from the Raw Peak Table of the peptide fragmentation spectrum. According to the method, the paired peak values (*m/z* and intensity) which were extracted from the raw spectrum according to the 'peak-picking' algorithm described above, are read into the

Raw Peak Table. The calculations are designed to eliminate values from ^{13}C isotopes (deisotoping), removal of a 2 Da precursor ion window and removal of values unlikely to be from y, a, or b ions. An input value for the mass resolution of the mass spectrometer used to obtain the spectrum is also required. The spectrum then undergoes scrutiny to eliminate certain peaks from the peak table. First, peaks are identified which are ~1 Da apart, where the higher mass peak is interpreted as arising from the presence of ions containing ^{13}C isotopes substituted for the more commonly occurring ^{12}C. For peaks identified as spaced 1 Da apart, an embodiment of the method eliminates the peaks of higher *m/z* from the peak table if the higher *m/z* peaks are of lower intensity than that of the peak 1 Da below. Second, the precursor ion is removed from the peak table. For a doubly charged precursor ion a window of *m/z* 0.5 below the precursor ion *m/z* and 1.5 above the precursor ion *m/z* is calculated and this region is removed. Any ions defined in the peak table which fall within this mass range are removed from the peak table. In a preferred embodiment to the present invention, a procedure of intensity thresholding is then carried out. This has the effect of removing peaks in the spectrum with a sufficiently low intensity value that they are considered unlikely to arise from the main fragment ion types relevant to the spectral interpretation i.e. y-ions, b-ions and a-ions. Without a degree of intensity thresholding the number of possible spectral interpretations significantly increases, hence slowing down the computational process.

A preferred embodiment of the thresholding procedure is as follows: equation (1) is used to compute the integer "ideal" number of peaks (P_{ideal}) which we would expect in a spectrum with a given precursor ion mass, *M*.

$$P_{Ideal} = 3 \times \left(\left(\frac{M}{100} \right) + 1 \right) \qquad (6)$$

The physical justification for using this equation is that we expect on average (*M*/100)+1 peaks due to the primary ion series (the y-ions), since the mean molecular weight of an amino acid residue is about 100 Da. Since three types of sequence ions are considered (y, b, and a ions) a multiplier of 3 is placed in equation (6). In reality there will be more peaks present due to internal fragmentation of the peptide and decay of ions into additional alternative decay products. The initial intensity threshold is then set to a value corresponding to the addition of 1 to the intensity of the lowest intensity peak in the

spectrum (the intensity refers to the total ion current). The thresholding algorithm then calculates the number of peaks that would remain in the spectrum if all peaks of intensity less than or equal to this threshold level are removed. If this number of peaks is greater than the ideal number computed in equation 1, then the intensity value is incremented by one and the calculation repeated. This procedure continues until the number of peaks remaining after application of the intensity threshold is less than or equal to the ideal number. This value is then decremented by one. The purpose of this is to produce a spectrum which has a larger number of peaks than the "ideal" number, but where most of the low intensity peaks have been removed. This threshold level is then applied to the peaks in the peak table object, and those peaks with intensity less than or equal to the threshold value are excluded.

In one embodiment of the invention, the sequencing loop of the program is then invoked. A description of the steps involved in the process is shown in the flow chart in **FIG.** 5. A select number of the highest *m/z* values remaining in the edited Peak Table object (typically twenty are selected) are then used to create Walk Objects and are placed on the Stack. In this process we assume that the set of the twenty starting *m/z* walk objects selected above the doubly charged ion will include y ions, and the walking process for each ion is carried out on the basis that we are starting with a y ion and walking down to a lower *m/z* y ion. We then calculate the mass difference between two peaks in the spectrum as though they were consecutive y ions in order to determine whether that mass difference could correspond to the mass of an amino acid, or modified amino acid. This value is determined within the cumulative error defined by the errors associated with each individual peak (in one embodiment the mass range of this error for the analysis to be equal to [(mass resolution)*$\sqrt{2}$]). Therefore, in the method following, the *m/z* values present in a particular walk object describe the mass-to-charge ratios of the y-ion fragments formed from the precursor ion.

In one embodiment, the first walk object from the stack is copied into a Walk Object known as the 'Current Walk'. The 'walk down' stage then proceeds. The lowest *m/z* value in the Current Walk, M, is determined (for the very first walk this will be the starting *m/z* value). The program then searches through the peak table for all *m/z* values lower than this value and tests whether the difference between the two *m/z* values corresponds to any of the amino acid residue masses defined in the amino acid object. In this process, the two *m/z* ions spaced apart by the mass difference corresponding to an

amino acid residue are assumed to be two consecutive y ions. If this is the case, then this value is a possible amino acid to add to the sequence defined in the Current Walk. As there may be more than one possibility for a correspondence, the program tracks the number of possible permutations. If there is only one possibility, then the *m/z* value of this possibility is added to the current walk, and the updated current walk is then resubmitted to the 'walk down' stage for further processing. If there is more than one possibility, the Current Walk is cloned as many times as necessary, and the appropriate *m/z* and sequence information is added to the clones, and these clones are added to the end of the stack object. If no possibilities exist, then the Current Walk has terminated at that position, and is copied back into its original position in the Stack. The stack counting index, which refers to the position of the Current Walk, is incremented. If unfinished walk objects remain on the stack, then the next incomplete walk object is taken from the Stack and becomes the new Current Walk. The process continues until there are no more incomplete Walk Objects in the Stack array. In one embodiment of the main sequencing process, all sequences possibilities generated from the 20 initial starting walk objects are kept within the stack, and there is no elimination or pruning process.

At this stage, Walk Objects in the Stack are reviewed, and an embodiment of the process is illustrated in the flow chart in **FIG.** 6. Firstly, sequence solutions are compared against each other to eliminate duplicate solutions. Sequence solutions are also rejected where the core sequence contains an internal lysine or arginine amino acid unless the amino acid immediately following at the carboxyl terminus side is a proline. Furthermore, in an embodiment of the method, the *m/z* values present in a particular walk object describe the mass-to-charge ratios of the y-ion fragments formed from the precursor ion. There is always the possibility, however, where b ions could have been incorrectly classified as y ions, and hence where amino acid sequences within the list of sequence solutions have been determined using b ions. This situation arises since there has been no pre determination of ion identity or of the originating terminus of a sequence ion, and all ions calculated to be above the intensity threshold (the procedure for which is described above) are used in the 'walking' or sequencing process. In order to eliminate possibilities where b ions have been incorrectly used instead of y ions amino acid sequences, the following procedure is used. Firstly the complementary ions to all *m/z* values in the Walk Object (which are assumed to be y ions) are calculated. For each y-ion formed from a doubly charged precursor ion, a complementary b-ion must also be

formed, which corresponds to the remainder of the precursor ion after fragmentation. (Biemann, **Anal. Chem.** 58, 1288A-1300A (1986), the contents of which is hereby incorporated by reference in its entirety) In the majority of cases, this b ion would be expected to be present in the spectrum, unless it has undergone further decomposition. The mass of this ion can be computed from the mass of the y-ion and the precursor ion mass. In one embodiment of the present method, the HOPS program calculates the *m/z* of all complementary ions to those present in the walk object. The *m/z* values for these complementary ions are then compared against the *m/z* values of the ions in the Walk Object. In any case where the *m/z* of the complementary ions is also present as an *m/z* value in the Walk Object then this sequence is rejected. The basis for this elimination process is the assumption that where a b ion has been incorrectly assumed to be a y ion in the sequencing process, there exists an *m/z* entry in the Walk Object for the complementary y ion to this b ion. Following the process of review, a list of all possible sequences that can be derived from the fragmentation spectrum is produced. The next task is to classify which of these is most likely to be the correct one through ranking and consensus procedure.

FIG. 7 shows an embodiment of the procedure implemented in the method to order the multiple walk object sequences. The first method of ranking is based on the total number of y, b and a ions in the fragmentation spectra which match the walk object sequences (ion-number ranking). The second method of ordering uses the sum of individual ion intensities which correspond to the signals assigned to y, b or a ions in the in the fragmentation spectra. A deduced sequence which has a higher value of summed ion intensities is considered more likely than those of lower summed intensities. In a preferred embodiment, the present invention sorts both by ion number rank and then by ion intensity rank and uses both top-ranked sequences (if different) for construction of the search string.

FIG. 8 shows an embodiment of the steps used to create a set of candidate search sequences which are derived from a y ion correlation analysis of the sequences returned from the ranking method (**FIG.** 7). The consensus sequence is defined as the common partial sequence within the set deduced for each fragmentation spectra. The consensus sequences are determined by calculating the frequency of occurrence of each y ion signal across the set of top ranked Walk Objects. Walk Objects which are supported by y ion signals for each residue are placed into the consensus set. An amino acid consensus

sequence is produced based on mass differences between the common set of sequential y ions which are related by the masses of the 20 naturally-occurring amino acids. In one embodiment, if the output of the ranking and y-ion correlation process yield a single sequence, it is used as the sole amino acid sequence for the candidate sequence set. If the consensus sequences are not the same length, in one embodiment the longest one is selected and truncated according to rules for constructing a 'hook' search string for the database of interest. For sequences of the same length, in one embodiment preference is given to those deduced from y ions with m/z values that are of greater value than the m/z value of the doubly-charged precursor ion.

It is well-recognised by those skilled in the art that spectra from different instruments and samples prepared by diverse methods vary in the levels of instrument and chemical noise, both across the full *m/z* range (upper limit is defined by the mass of the precursor ion) and within defined *m/z* regions. Further, contamination of the fragmentation spectrum with ions other than those of the precursor peptide ion of interest may contribute significantly to the ions observed at m/z values below the doubly charged precursor ion. As shown in **FIG. 8**, the HOPS method can use multiple criteria to select sequences for candidate sequence selection to compensate for differences among samples and instruments. For example, using the y ions above the doubly charged precursor ion for certain fragmentation spectra may increase the fidelity of spectral sequence reads since contaminating fragment ions from singly-charged, unrelated peptides having the same mass (± 2 Da) as the precursor ions would not be present. We have determined from studies using known peptides that the HOPS process produces a consensus sequence containing the correct trimer sequence in 94.5% of the computer interpreted spectra, without any user qualifications. In one embodiment of the present invention, all consensus sequences are used to produce mass-sequence-mass search strings or 'hooks' to search databases containing translated genomic sequences.

In one embodiment, the candidate search sequences, which are the 'high fidelity' output of HOPS are then passed to the FIREPEP module (**FIG. 9**). The initial input are the HOPS candidate search sequences deduced from the set of fragmentation spectra for each protein. In one embodiment, the first step is to determine which of the candidate sequences is suitable for constructing the 'hook' search string. In this embodiment, two considerations are applied, *i.e.* the number and the composition of amino acid residues to be included in the 'hook' search string. In a preferred embodiment for the six-frame

translated human genome (~3 x 10^9 nucleotides), the candidate search sequences must contain at least three amino acid residues. Search strings consisting of dimers have utility for searching smaller genomes such as those from microorganisms (Mann and Wilm (1994) Anal Chem. 66, 4390). However, when the output from HOPS is used to identify proteins encoding genes in larger genome databases, a trimer sequence flanked by two masses is required to retrieve a practical working set of translated nucleotide sequences from a large genomic database. The HOPS method can construct search strings with dimers or sequences greater than three amino acid residues in length but, in a preferred embodiment, the additional sequence length may be necessary and may increase the number of false positive reads from the fragmentation spectra, thus compromising the fidelity of the overall process. However, it should be noted that in the present invention, any sequence data that do not meet the criteria required to form a 'hook' search sequence may be used to search the retrieved protein or translated nucleotide database sequences and the false positives removed using a spectral 'back read' algorithm (**FIG.** 2, step 7 and **FIG.** 11).

In one embodiment, the second step is to define the amino acid composition of the sequence in the 'hook' search string. The criteria to enable specific searching of the human genome is shown in **FIG.** 10 under 'Constraints for search string-trimer'. These are based on empirical observations and from database attributes. One of ordinary skill in the art would readily recognise that criteria applicable to other genomes could be derived in a like manner or theoretical models which encompass such parameters as genome size, frequency of translated amino acid residues, nucleotide sequencing error rates and gene structure. Other considerations for trimer composition are related to (a) the elemental identity of certain residues which, therefore, have identical mass, and (b) instrument performance which produces sequence ambiguities for both single and multiple residues. These alternative embodiments fall within the method described herein. To account for these mass identities and similarities among the naturally-occurring amino acid residue the following may be considered:

1. The amino acids leucine (L) and isoleucine (I) are isobaric isomers *i.e.* these residues have identical chemical composition (Table 1). Here, the HOPS algorithm uses the symbol L, and this may be permuted to I to form the single residue change. For

construction of the allowed permuted sequence set (**FIG.** 9), 'hook' sequences with both Leu and Ile are used.

2. The amino acid residue, phenylalanine (F) has a mass which may be similar to the oxidised form of methionine (M*) (147.0684 versus 147.0399). Depending on signal intensity and instrument resolution, it may be difficult to achieve sufficient mass accuracy to distinguish these two residues. In a preferred embodiment, the HOPS algorithm always specifies F for both F and oxidized methionine (M*). Here, allowed 'hook' sequence construction, peptides with both F and M* are included.

3. The amino acid residue, glutamine (Q) has a mass which may be similar to lysine (K) (128.0586 versus 128.0950). In a preferred embodiment, the HOPS algorithm uses the Q with a change to K for the single residue change. In building the allowed trimer sequences for tryptic digests, K is included only if followed by a Pro residue which is know to be resistant to trypsin cleavage.

4. The amino acid tryptophan (W) has a mass which may be similar to the masses of three amino acid residue dimers and produces six sequence permutations, as shown below. This case occurs when a signal between the two defining the tryptophan residue (mass = 186.0793) in the fragmentation spectrum is below detectable limits.

(i) alanine (A) and aspartic Acid (D), either AD or DA (mass difference = 0.015256 Da) (ii) glycine (G) and glutamic Acid (E), either GE or EG (mass difference = 0.015256 Da) (iii) valine (V) and serine (S), either VS or SV (mass difference = 0.21129 Da). Here, cases where the 'hook' search string contains W, a set of six permuted search strings could be constructed. Alternately, if the candidate search sequence contains one of the 6 di-amino acid combinations, the two residues could be permuted to a Trp ('back permutation').

5. The amino acid asparagine (N) and two glycine residues (GG) are isobaric isomers and have identical chemical compositions and hence mass.

6. The amino acid arginine (R) (156.1011) has a similar mass to a combination of valine (V) and glycine (G), either VG or GV (mass difference = 0.0011232 Da).

All permutations are cumulative, and so all possible combinations must be calculated from the candidate sequences to generate the 'Permuted Search Set' of text

strings. For example, a partial amino acid sequence of LCW would generate the following 14 possibilities in the Permuted Search set: LCW, ICW, LCDA, ICDA, LCAD, ICAD, LCGE, ICGE, LCEG, ICEG, LCVS, ICVS, LCSV, ICSV.

In this embodiment of the present invention, the first three sequence permutations (L↔I, F↔M* and Q↔K) are used in the construction of the search string 'hooks' in the example of the present invention. The remaining permutations may have utility for other translated genome and protein databases. In one embodiment of the present invention permutations from one amino acid to multiple amino acids (*e.g.* N↔GG and W↔(AD or DA or VS or SV or EG or GE)) are not considered since only sequences that are supported with observed peaks in fragmentation mass spectra are used for construction of the 'hook' search string.

In one embodiment, after selection of the subset of candidate sequences, which contain trimers with allowed combinations of amino acids, the 'hook' search string is completed by calculating the first and second masses, M1 and M2 which are the masses from the start of the trimer sequence to the N terminus of the peptides and the end of the trimer sequence to the C terminus of the peptide, respectively. These values are determined within the cumulative error defined by the errors associated with each individual peak (we have calculated the mass range of this error for our current analysis to be equal to [(mass resolution)*√2]). M1 is calculated as the difference between the mass of the singly-protonated ion (determined by MALDI-TOF MS of the mixture of peptides for the protein) and the higher *m/z* value which defines the bounds of the trimer in the original fragmentation spectrum. For a sequential determination of product ions assumed to be y ions, M1 is equal to the sum of the amino acid residues which extend from the N-terminal of the database peptide sequence to first trimer residue. The C-terminal mass or M2 is the value of the lowest *m/z* value which defines the trimer sequence. For a sequential determination of product ions assumed to be y ions, M2 consists of the masses for the unknown amino acid residues, the mass of a molecule of water and the proton added from the ionisation process. Thus, the sum of M1, residue masses of the trimer and M2 equal the molecular weight of the singly protonated molecular ion of the peptide, which corresponds to the observed value from MALDI/MS analysis within the defined mass tolerance. To allow for the search string constraint that M1 cannot equal the mass of any of the naturally-occurring amino acid residues, the first residue of the trimer sequence must be at least the third residue from the N-terminal of

the peptide (**FIG.** 10). In the method of the present invention which utilises trypsin as the method of specific cleavage for generation of peptides, the last residue of the peptide cannot be the C-terminal Arg or Lys residue. Therefore, in order to satisfy the M2 constraint the mass of M2 must be greater than the combined mass of an arginine residue with a water molecule and proton added (> 175.12 Da) (**FIG.** 10).

In the preferred implementation, the 'hook' amino acid sequence is deduced from the spectrum assuming that sequence order is determined by a y ion series. If the ions considered in the sequential determination are in fact b ions, the sequence orientation would be reversed (**FIG.** 1) and the values of M1 and M2 would change accordingly. For example, a string produced with a y ion spectral read, (NH_2)-M_1-Leu-Val-Ala-M_2-(COOH), would become (HOOC)-M_2-Ala-Val-Leu-M_1-(NH_2), if b ions are used to deduce the sequence. Both difference in sequence orientations and flanking mass values can be considered in the FIREPEP search algorithms described below.

The detailed rules of the preferred embodiment for the construction of a 'hook' search string with practical specificity for retrieving a small set of translated nucleotide sequences from the completed human genome are summarised in **FIG.** 10 and are as follows:

i) the mass of M1 cannot equal the mass of a single naturally-occurring amino acid residue.

ii) the mass of M2 must be greater than the mass of a protonated arginyl amino acid

iii) only single permuted residues (L→I, F→M*, or , Q→K if the K is followed by a P residue) is allowed within the trimer;

iv) the trimer sequence cannot contain only combinations of the high frequency residues V or A or combinations of either V or A residue and a single permutated residue. For example, IVA would not be allowed as the trimer sequence of a 'hook' search string.

v) permuted sequences which are based on mass similarities between the mass of a single amino acid residue and residue dimers are not incorporated into the search string 'hook'.

One of ordinary skill would readily recognize alternative criteria for retrieval of the correct nucleotide sequence, and such criteria may be readily incorporated into the method.

In a preferred embodiment, the FIREPEP module uses the 'hook' M1-trimer-M2 strings to search a stored *in silico* generated database of endoprotease-defined peptides

produced from protein sequences and translated genomic sequences. In one example of the present invention, a database was constructed from a combination of human genomic sequence entries in the database held at the European Molecular Biology Laboratory (EMBL), protein entries in the non-redundant database held by the National Centre for Biotechnology Information (NCBI) which is accessible at http://www.ncbi.nlm.nih.gov/ and protein sequence entries held at the SWISSPROT database held at the Swiss Institute of Bioinformatics (SIB). The genomic sequences are incorporated into the peptide database by first translating, bidirectional in all three frames, all nucleotide sequences, edited as well as those unassembled, unordered segments of a genome. The resulting amino acid sequences are cleaved at all residues predicted by the specificity of the endoprotease of interest. In the example of *in silico* trypsin cleavage, Lys and Arg cleavage sites are defined which results in peptides that do not contain more than one Lys or Arg residue, except those with a putative C-terminal Pro residue. Peptides containing stop codon nucleotide sequences (usually indicated by an '*' in the translated database) or characters not corresponding to the 20 amino acids (Table I) are not entered into the peptide database. Peptides with translated nearest neighbor stop codons (one or two flanking residues) are rejected. Both the oxidized and unoxidized form of peptides containing Met, as well as all permutations if the peptide contains multiple Met residues, are entered into the database. The rules used to build the peptide database satisfy the 'Allowed Database Sequence' constraints detailed in **FIG.** 10. In summary, the combination of criteria for building the 'hook' search string and those used to build the peptide database confer a high specificity to the FIREPEP search engine.

In a preferred but certainly not the only embodiment, the FIREPEP method (shown in **FIG.** 9) then takes the set of 'hook' search strings and performs a text search for the trimer sequence in database peptides with masses corresponding to the mass of the 'hook' peptide string (singly-protonated molecular weight of the peptide of interest). If at least one of the 'hook' trimer peptide sequences is found in this mass-constrained subset of the peptide digest database, FIREPEP then carries out mass reconciliation calculations between the 'hit' peptides and the M1 and M2 values. The M1 is mass-matched to the characters immediately preceding the 'text-matched sequence' *i.e.* toward the putative N-terminal. For M1, the program sequentially calculates the mass of the amino acid residues represented by the database characters starting at the residue closest to the 'N-terminal' end of the 'text matched sequence'. The process continues until the

total mass of the character set exceeds the value of M1. If the value of the character set is equal to the value specified by M1 (within pre-determined error ranges dictated by the mass spectrometer resolution as described above), then an N-terminal mass match is made and the sequence is 'passed' for C-terminal mass matching. A similar procedure is then carried out starting at the C-terminal end of the 'text-matched sequence' for the C-terminal mass value. In this case, the mass of 18.0153 Da (from a water molecule associated with the y ion and 1.0079 Da for the mass of the ionising proton adduct) must be added to reconcile the mass-sequence extension on the C-terminal side of the string. If the adjusted masses of the residues is equal to the experimentally determined C-terminal mass, then the mass-matching process is deemed successful. If either of the two mass matching criteria are not met, then the search fails, and the program moves to the next peptide sequence for which there is a 'text-matched sequence'. The output of the FIREPEP search is one or a small set of peptides related by total mass, trimer sequence and mass-matched sequences for M1 and M2. In cases where there are multiple peptide sequences returned, the correct sequence is determined by a spectral 'back read' (y, b and a ions are considered) of each candidate peptide sequence,. In the present invention, this spectral-sequence comparison proceeds in a user-independent mode. The resulting correct peptide sequence is then used to call (via cross-indexing routines) the full-length translated genome sequences from the genome database. At this stage, the specificity of the process is defined as one fragmentation spectrum yields a single database sequence.

Translated nucleotide databases as amino acid sequences usually contain both genome and mRNA sequences of varying lengths and coverage of the expressed protein products. In many cases the retrieved database sequences do not contain the full-length sequence which is found in the biologically translated protein. Since in the present invention, multiple fragmentation spectra and accurate peptide molecular weights are acquired for a single protein, the multiple peptide sequences and masses have utility for 1) identifying expressed regions or exons within nucleotide sequences; 2) setting the correct nucleotide reading frame within the expressed regions or exons; 3) clustering related sequences which have not been assemble or assigned to a gene family; 4) identifying database redundancy 5) identifying sequencing artefacts and errors in nucleotide sequences and 6) identifying base change mutations such as those observed in polymorphisms. In one embodiment, the consolidation of 'hit' database sequences using tandem sequences as anchors is performed by the Find Related Protein module

(FIREPPROT) (**FIG.** 11). The tandem sequence data is from the one or more 'hook' peptide sequences from the output of FIREPEP and all HOPS sequences that did not meet the criteria of the Consensus Procedure (**FIG.** 8). As shown in **FIG.** 11, FIREPROT reads in all HOPS data, including both consensus and unedited sequences and creates a permuted search set which, unlike the constraints required for the specificity of 'hook' sequences, considers sequence permutations from mass ambiguities. For example, the six possible sequence permutations due to the mass ambiguity associated with a Trp residue are calculated and placed into this 'permuted sequence set'. In addition, peptides from incomplete tryptic cleavages and those with multiple flanking K and R residues are allowed. First a text search of a HOPS sequence is made for all hooked database sequences. If a text match is found and M1 and M2 mass matching criteria are met, then the HOPS sequence is mapped onto one or more of the retrieved database sequences. In this way all tandem sequence data, including dimer sequences can be mapped onto the database sequences retrieved by the 'hook' sequence strings.

In one example, a combined translated human genome and protein database (150 million peptides) was searched with two 'hook' M1-sequence-M2 strings, namely 226.15-NEN-621.34 and 276.15-WDD-260.20 constructed by HOPS from tandem mass spectral analysis of two different peptides from a digest mixture of a protein. **FIG.** 12 shows the alignment and clustering of the human genome sequences for the transferrin receptor identified using these two 'hook' sequence anchors, and the results of mapping additional HOPS consensus sequence data *i.e.* the M1-dimer-M2 718.37-SF-274.19. The HOPS sequences are underlined. For purposes of presentation, the genome sequence containing the 'hook' sequence has been extended only to the flanking translated stop codons. Genome sequences which do not agree with the protein template can be easily identified from the alignment.

In addition to mapping all HOPS-derived sequences from tandem spectra to the consolidated set of translated genomic sequences, mass-mapping software is used to assign observed masses to peptide sequences in the genome sequences. In addition, observed masses are attributed to molecular weight increments and decrements.to account for post-translational modifications of interest. As shown in **FIG.** 13, the output of FIREPROT is read into the mass-mapping and post-translational modification module. The first step is to produce all possible peptide masses for the 'hit' sequences and masses for the post-translational modifications of interest. For example, to consider

phosphorylation, the mass of all peptides containing Ser, Thr or Tyr residues are incremented by a mass = 79.9663 for each of the residues which could be phosphorylated. All possible combinations are calculated. Thus, the modified peptide mass list consists of all masses, unmodified and modified, for each peptide which could result from endoprotease digestion and post-translational modifications of interest. The modified mass list is then used to compare with the experimentally-determined peptide mass list. All mass agreements with translated genome peptide sequences within the mass tolerance of the instrument are used to determine sequence coverage and post-translational modifications. Table II lists the masses assigned to peptides unmodified and modified to the transferrin receptor gene. **FIG.** 14 summarises the mapping of observed peptide masses which match the molecular weight of the translated genome peptides.

It is well recognised that changes in single nucleotides occur with high frequency among the genomes of individuals. The biological consequences of single nucleotide polymorphisms (SNPs) can be profound. The propensity to develop a form of Alzheimer;s disease is associated with a single base pair change in apolipoprotein E (Apo E) which converts the Cys residues at position 130 (the ε3 isoform) to Arg (ε4 isoform). The tryptic peptides containing the amino acid change from the SNP are (R)LGADMEDVCGR and (R)LGADMEDVR for isoforms ε3 and ε4, respectively. Apolipoprotein E peptides from individuals with the ε3 and ε4 polymorphism were analysed according to the described invention. Two SNP specific tryptic peptides were produced which yielded the following HOPS consensus sequences: M1=241 (DME) M2=573 (ε3), and M1=241 (DME) M2=476 (ε4). There SNP hooks were used to search the human genome peptide database with the FIREPEP module. In the case of the ε3 search string Apo E gene sequences were returned. The search with ε4 did not return any records. These results indicate that ApoE DNA sequences from individuals with the ε4 SNP are not in the current publication of the human genome.

Table 1. Amino acid residue monoisotopic masses

Amino Acid	Symbol	Elemental Composition	Monoisotopic mass (Da)
Alanine	A	C_3H_5NO	71.037114
Arginine	R	$C_6H_{12}N_4O$	156.10111
Asparagine	N	$C_4H_6N_2O_2$	114.042927
Aspartic Acid	D	$C_4H_5NO_3$	115.026943
Carboxyamido Cysteine[1]	C	$C_5H_8N_2O_2S$	160.03065
Glutamic Acid	E	$C_5H_7NO_3$	129.042593
Glutamine	Q	$C_3H_8N_2O_2$	128.058577
Glycine	G	C_2H_3NO	57.021464
Histadine	H	$C_6H_7N_3O$	137.058912
Isoleucine	I	$C_6H_{11}NO$	113.084064
Leucine	L	$C_6H_{11}NO$	113.084064
Lysine	K	$C_6H_{12}N_2O$	128.094963
Methionine	M	C_5H_9NOS	131.040485
Oxidised Methionine	M*	$C_5H_9NO_2S$	147.035340
Phenylalanine	F	C_9H_9NO	147.068414
Proline	P	C_5H_7NO	97.052764
Serine	S	$C_3H_5NO_2$	87.032028
Threonine	T	$C_4H_7NO_2$	101.047678
Tryptophan	W	$C_{11}H_{10}N_2O$	186.079313
Tyrosine	Y	$C_9H_9NO_2$	163.063328
Valine	V	C_5H_9NO	99.068414

[1]*All Cysteines are modified to the carboxyamino derivative during our production process.*

Table 2. Assigned masses for modified and unmodified peptides to the Transferrin receptor gene

Mass of singly protonated peptide	Maldi Peptide Matches		ppm
1084.59209	AFTYINLDK		-22
936.48636	AVLGTSNFK		31
872.43325	DAWGPGAAK		-8
1610.80827	DENLALYVENQFR		-17
806.39179	DGFQPSR		-15
808.40533	DLNQYR		-13
773.39502	DQHFVK		-1
1288.68379	DSAQNSVIIVDK		-7
1217.56546	EEPGEDFPAAR		-18
1672.8441	GFVEPDHYVVVGAQR		0
1977.92353	HPVTGQFLYQDSNWASK		11
713.41117	IPELNK		12
708.38883	ITFAEK		6
1282.64946	LAQMFSDMVLK		3
1561.72694	LAVDEEENADNNTK		-19
1197.61076	LDSTDFTGTIK		-9
1655.729	LFGNMEGDCPSDWK		-28
1204.65388	LLNENSYVPR		-17
1616.8125	LTHDVELNLDYER		-13
1095.51951	LTTDFGNAEK		11
1033.53257	LTVSNVLK		-57
986.52812	LVHANFGTK		14
952.48181	LYWDDLK		-5
797.42723	MVTSESK		-69
1358.61735	QVDGDNSHVEMK		-12
1745.87097	SAFSNLFGGEPLSYTR		-12
958.51392	SGVGTALLLK		83
1468.84215	SSGLPNIPVQTISR		-20
1565.80219	VEYHFLSPYVSPK		-1
1513.80462	VSASPLLYTLIEK		-8
1226.66333	YNSQLLSFVR		-8
1344.690	RLY*WDDLKR	(+1P)	-33
1316.670	LY*WDDLKRK	(+1P)	-24
788.398	IT*FAEK	(+1P)	-48
1513.805	VS*AS*PLLY*T*LIEK	(+1P)	-8

**indicates a phosphorylated residue. The number of phosphorylated sites is indicated in brackets next to the sequence. Where more than one residue in a peptide is given an asterisk, the phosphorylated residue could be either one, or a combination, of the residues marked with an asterisk to give the total number of phosphate groups identified for that peptide mass.*

WE CLAIM:

1. A method for determining an amino acid sequence of a peptide which comprises:

(a) obtaining a suitable fragmentation mass spectrum having a plurality of peaks for the peptide;

(b) removing the peaks due to C13 isotopes from the spectrum and applying an appropriate intensity threshold to the remaining peaks;

(c) selecting a suitable peak as a starting point and determining mass differences in the spectrum that corresponds to an amino acid residue mass difference;

(d) sequentially determining each subsequent amino acid residue mass difference from the starting peak;

(e) repeating the process of steps (c) and (d) for additional peaks so as to obtain a non-redundant set of proposed sequences; and

(f) ranking the proposed sequences and preparing a consensus sequence corresponding to the amino acid sequence of the peptide.

2. The method of claim 1 wherein the peptide is obtained by selective cleavage of a polypeptide.

3. The method of claim 1 wherein the peptide is obtained by selective cleavage of a polypeptide and any proposed amino acid sequences inconsistent with the selective cleavage to are removed obtain a revised set of proposed sequences prior to ranking the proposed sequences.

4. The method of claim 1 wherein the suitable peak as the starting point is a high mass peak.

5. The method of claim 1 wherein the sequential determination of step (d) is repeated until nearly every possible amino acid residue mass difference is investigated.

6. The method of claim 1 wherein the sequential determination of step (d) is repeated until every possible amino acid residue mass difference is investigated.

7. The method of claim 1 wherein the amino acid sequence is a partial amino acid sequence.

8. The method of claim 7 wherein the partial amino acid sequence is a trimer.

9. The method of claim 1 wherein the amino acid sequence is a complete amino acid sequence.

10. The method of claim 2 wherein the selective cleavage is enzymatic cleavage.

11. The method of claim 10 wherein the selective enzymatic cleavage is cleavage with arginine endopeptidase (ArgC); asparatic acid endopeptidase N (aspN); chymotrypsin; glutamic acid endopeptidase C (gluC); lysine endopeptidase C (lysC); trypsin; or V8 endopeptidase.

12. The method of claim 1 wherein the suitable fragmentation mass spectrum is obtained from a suitable mass spectrometer.

13. The method of claim 12 wherein the suitable mass spectrum is obtained from a tandem mass spectrometer.

14. The method of claim 13 wherein the tandem mass spectrometer is a quadropole tandem time of flight mass spectrometer (Q-TOF).

15. The method of claim 1 wherein the sequential determination of the amino acid sequence is a sequential determination of a y-ion series.

16. The method of claim 1 wherein the sequential determination of the amino acid sequence is a sequential determination of a b-ion series.

17. The method of claim 1 wherein about five to about fifty peaks are selected in step (e) as additional peaks for the sequential determination of the amino acid mass residue difference.

18. The method of claim 1 wherein about ten to about thirty peaks are selected in step (e) as additional peaks for the sequential determination of the amino acid mass residue difference.

19. The method of claim 1 wherein the suitable fragmentation mass spectrum has a minimum resolution of at least 5600 full width half peak height for peptides with a molecular weight up to about 4000 daltons.

20. The method of claim 1 wherein the suitable fragmentation mass spectrum has a minimum resolution of at least 2800 full width half peak height for peptides with a molecular weight up to about 2000.

21. The method of claim 20 wherein the suitable fragmentation mass spectrum has a minimum resolution of at least 4000 full width half peak height.

22. The method of claim 1 wherein the peptide comprises at least one post-translationally modified amino acid.

23. The method of claim 1 wherein the peptide comprises a plurality of post-translationally modified amino acids.

24. The method of claim 1 wherein the polypeptide is a polypeptide that has not been disclosed in a publically available database.

25. The method of claim 1 wherein the determined amino acid sequence is compared to a conceptually translated nucleotide sequence in a database and the results of the comparison are used to identify expressed regions in the nucleotide sequence.

26. The method of claim 7 wherein the partial amino acid sequence is combined with a total mass of the peptide to obtain a first mass and a second mass for the peptide and utilizing the first mass, sequence second mass combination to identify a polypeptide in a database.

27. The method of claim 26 wherein the polypeptide in the database is a conceptually translated polypeptide from a nucleotide database.

28. The method of claim 26 wherein the polypeptide in the database is a polypeptide from a protein database.

29. The method of claim 26 wherein in the database is a combined database including a conceptually translated polypeptides from a nucleotide database and polypeptides from a protein database.

30. The method of claim 26 wherein neither the first mass nor the second mass is a single amino acid residue mass.

31. The method of claim 26 wherein the partial amino acid sequence is a trimer.

32. The method of claim 26 wherein the database contains more than about 200,000 sequences.

33. The method of claim 26 wherein the database contains more than about 500,000 sequences.

34. The method of claim 26 wherein the total mass is determined to an error of measurement of about 200 parts per million (ppm) or less.

35. The method of claim 30 wherein the total mass is obtained from a matrix assisted laser desorption time-of-flight (MALDI-TOF) mass spectrometer.

36. The method of claim 25 or 26 wherein the database is a database of sequences from the human genome.

37. The method of claim 25 or 26 wherein the expressed regions or identification are used to determine an error due to a nucleotide mutation.

38. The method of claim 25 or 26 wherein the expressed regions or identification are used to determine an error due to a nucleotide sequencing.

39. The method of claim 25 or 26 wherein the method is repeated for a plurality of determined amino acid sequences.

40. The method of claim 25 or 26 wherein the method is repeated for a plurality of consensus amino acid sequences.

41. A method of identifying a polypeptide(s) in a database based on a peptide obtained by selective cleavage of an unknown polypeptide which comprises:
(a) obtaining a suitable first mass of two or more amino acids, a sequence of three amino acids selected such that only one amino acid is ambiguous due to isobaric amino acids and amino acids ambiguous due to the resolution of the instrument, and a suitable second mass of two or more amino acids from the peptide;
(b) prepare a permuted search set based on the suitable amino acid sequence by including possible sequences due to isobaric amino acids and amino acids ambiguous due to the resolution of the instrument;
(c) comparing the first mass, the permuted search set and the second mass to the database so as to obtain a set of possible polypeptides corresponding to the unknown polypeptide; and
(d) removing from the set of possible polypeptides any polypeptide inconsistent with the selective cleavage method so as to identify the polypeptide(s).

42. The method of claim 41 wherein the suitable sequence of three amino acids is chosen to exclude specific amino acid combinations that appear with a high frequency in the database.

43. The method of claim 41 wherein the database is a database of protein sequences.

44. The method of claim 41 wherein the database is a database of peptide sequences.

45. The method of claim 41 wherein the database is a database of peptide sequences prepared from a database of protein sequences.

46. The method of claim 41 wherein the database is a database of peptide sequences prepared from a database of nucleic acid sequences.

47. The method of claim 41 wherein the database is a combined database of peptide sequences prepared from a database of nucleic acid sequences and a database of protein sequences.

48. The method of claim 41 wherein the database is constrained to remove any sequences that include a residue that appears adjacent to a stop codon in the nucleotide database.

49. The method of claim 41 wherein the database contains over 200,000 sequences.

50. The method of claim 41 wherein the database contains over 500,000 sequences.

51. The method of claim 41 wherein the database contains over 1,000,000 sequences.

52. The method of claim 41 wherein the database contains over 10,000,000 sequences.

53. The method of claim 41 wherein the database contains over 100,000,000 sequences.

54. The method of claim 41 wherein the selective cleavage is enzymatic cleavage.

55. The method of claim 54 wherein the selective enzymatic cleavage is is cleavage with with arginine endopeptidase (ArgC); asparatic acid endopeptidase N (aspN); chymotrypsin; glutamic acid endopeptidase C (gluC); lysine endopeptidase C (lysC); trypsin; or V8 endopeptidase.

56. The method of claim 41 wherein the permuted search set is selected from one of the following permutations: arginine permuted with (valine and glycine) and vis versa; asparagine permuted with two glycines and vis versa; leucine permuted with isoleucine and vis versa; lysine permuted with glutamine and vis versa; phenylalanine permuted with oxidized methionine and vis versa; tryptophan permuted with (alanine and aspartic acid) and vis versa; tryptophan permuted with (glycine and glutamic acid) and vis versa; or tryptophan permuted with (valine and serine) and vis versa.

57. The method of claim 41 wherein the suitable amino acid sequence is obtained by

(e) obtaining a suitable fragmentation mass spectrum having a plurality of peaks for the peptide;

(f) removing the peaks due to C13 isotopes from the spectrum and applying an appropriate intensity threshold to the remaining peaks;

(g) selecting a suitable peak as a starting point and determining mass differences in the spectrum that corresponds to an amino acid residue mass difference;

(h) sequentially determining each subsequent amino acid residue mass difference from the starting peak;

(i) repeating the process of steps (g) and (h) for additional peaks so as to obtain a non-redundant set of proposed sequences; and

(j) ranking the proposed sequences and preparing a consensus sequence corresponding to the amino acid sequence of the peptide.

58. The method of claim 57 wherein the method is a method to determine which sequences are expressed in a nucleotide sequence database and the polypeptide has a conceptually translated sequence encoded by a nucleotide sequence;

(k) mapping the sequence of the peptide onto the nucleotide sequence; and

(l) mapping either the consensus or the proposed sequence for at least one additional peptide from the polypeptide on to the nucleotide sequence so as to map of portions the nucleotide sequence that are expressed.

59. The method of claim 58 wherein a plurality of additional peptides are mapped on to the nucleotide sequence.

60. The method of claim 57, 58 or 59 wherein a total mass for at least one additional peptide is mapped on to the nucleotide sequence.

61. The method of claim 57, 58 or 59 wherein a plurality of total masses for a plurality of additional peptides are mapped on to the nucleotide sequence.

62. The method of claim 57, 58 or 59 wherein the identification of the polypeptide is used to identify an expressed region in a nucleotide sequence.

63. The method of claim 57, 58 or 59 wherein the identification of the polypeptide is used to identify a protein isoforms.

64. The method of claim 57 wherein the suitable mass spectrum is obtained from a suitable mass spectrometer.

65. The method of claim 57 wherein the total mass is obtained from a matrix assisted laser desorption time-of-flight (MALDI-TOF) mass spectrometer.

66. A method of identifying polypeptides in a database based on a peptide obtained by selective cleavage of an unknown polypeptide which comprises:

(a) obtaining a suitable first mass, a suitable sequence of three or more amino acids and a second mass from the peptide;

(b) preparing a permuted search set based on the suitable amino acid sequence by including possible sequences due isobaric amino acids and amino acids ambiguous due to the resolution of the instrument;

(c) comparing the first mass, the permuted search set and the second mass to the peptide database so as to obtain a set of possible polypeptides corresponding to the unknown polypeptide; and

(d) removing from the set of possible polypeptides any polypeptide inconsistent with the selective cleavage method so as to identify the polypeptide.

67. A method of identifying polypeptides in a database based on a peptide obtained by selective cleavage of an unknown polypeptide which comprises:

(a) obtaining a suitable first mass, a suitable amino acid trimer and a second mass from the peptide;

(b) preparing a permuted search set based on the suitable amino acid sequence by including possible sequences due isobaric amino acids and amino acids ambiguous due to the resolution of the instrument;

(c) comparing the first mass, the permuted search set and the second mass to the database so as to obtain a set of possible polypeptides corresponding to the unknown polypeptide; and

(d) comparing a predicted mass spectrum for the possible peptides to an actual mass spectrum for the polypeptide so as to identify the polypeptide.

68. A method for identifying expressed regions in a nucleotide sequence present in a database which comprises:

(a) obtaining a peptide by selective cleavage of a polypeptide;

(b) obtaining a suitable first mass, a suitable sequence of three or more amino acids, and a second mass for the peptide;

(c) preparing a permuted search set based on the suitable amino acid sequence by including possible sequences due isobaric amino acids and amino acids ambiguous due to the resolution of the instrument;

(d) comparing the first mass, the permuted search set and the second mass to a database of polypeptides derived from conceptually translating the nucleotide sequences so as to obtain a set of possible polypeptides corresponding to the polypeptide; and

(e) removing from the set of possible polypeptides any polypeptide inconsistent with the selective cleavage method so as to identify the nucleotide sequence/(s) encoding the polypeptide in the database and the nucleotide sequence of the peptide encoded therein;

(f) repeating steps (a) through (e) for at least one more peptide obtained from the polypeptide;

(g) analyzing the results of steps (e) and (f) so as to identify expressed regions in a nucleotide sequence.

69. The method of claim 68 wherein the total mass for at least one additional peptide is mapped on to the nucleotide sequence.

70. The method of claim 68 wherein a plurality of total masses for a plurality of additional peptides are mapped on to the nucleotide sequence.

71. The method of claims 68, 69 or 70 wherein the total masses of the peptides are measured to an accuracy of 1ppm or less.

72. The method of claims 68, 69, 70 or 71 wherein the method is used to identify post-translational modifications present in the protein encoded by the nucleotide sequence.

73. The method of claim 68, 69, 70 or 71 wherein the method is used to identify single nucleotide polymorphisms.

74. The method of claim 68, 69, 70 or 71 wherein the expressed regions or identification are used to determine an error due to a nucleotide mutation.

75. The method of claim 68, 69, 70 or 71 wherein the expressed regions or identification are used to determine an error due to nucleotide sequencing.

76. The method of claim 68, 69, 70 or 71 wherein the expressed regions or identification are used to determine an intron or exon boundary.

77. The method of claim 1 wherein the amino acid sequence of the peptide is used to diagnose a disease.

78. The method of claim 65 wherein the amino acid sequence of the peptide is used to diagnose a disease associated with a specific post-translational modification.

ABSTRACT OF THE INVENTION

In one embodiment, the present invention describes a method for defining the physical DNA sequences in genomes which are expressed as proteins. In one embodiment, the invention consists of linked software modules to create an automated flow of data from primary mass spectral information to expressed gene sequences. The Holistic Peptide Sequencing (HOPS) module uses a novel spectral sequencing algorithm to produce sequence information with high accuracy from peptide fragmentation mass spectra. In another embodiment the Find Related Peptide (FIREPEP) software constructs search strings from selected HOPS sequences and the molecular weights of the respective peptides, and uses database specific criteria for allowed partial sequences to construct the database 'hook' string. The 'hook' string is then used to search a database of protein sequences and/or conceptually translated genome sequences, and/or peptide sequences. The result of the search is a single set of raw translated genome sequences and/or protein sequences which share the same HOPS peptide sequences found in the 'hook' search string. In yet another embodiment the Find Related Protein (FIREPROT) module consolidates the translated genome sequences and/or protein sequences by mapping all remaining HOPS sequences and observed peptide molecular weights onto the identified sequences. These methods may be used to organize intron and exon sequences into a coherent gene structure, correct artifactual nucleotide sequencing errors, define expressed genetic mutations and protein polymorphisms, characterise post-translational proteolytic processing and define the type, and the residue location of amino acid residue modifications.

FIG 1. Peptide Sequencing Ions

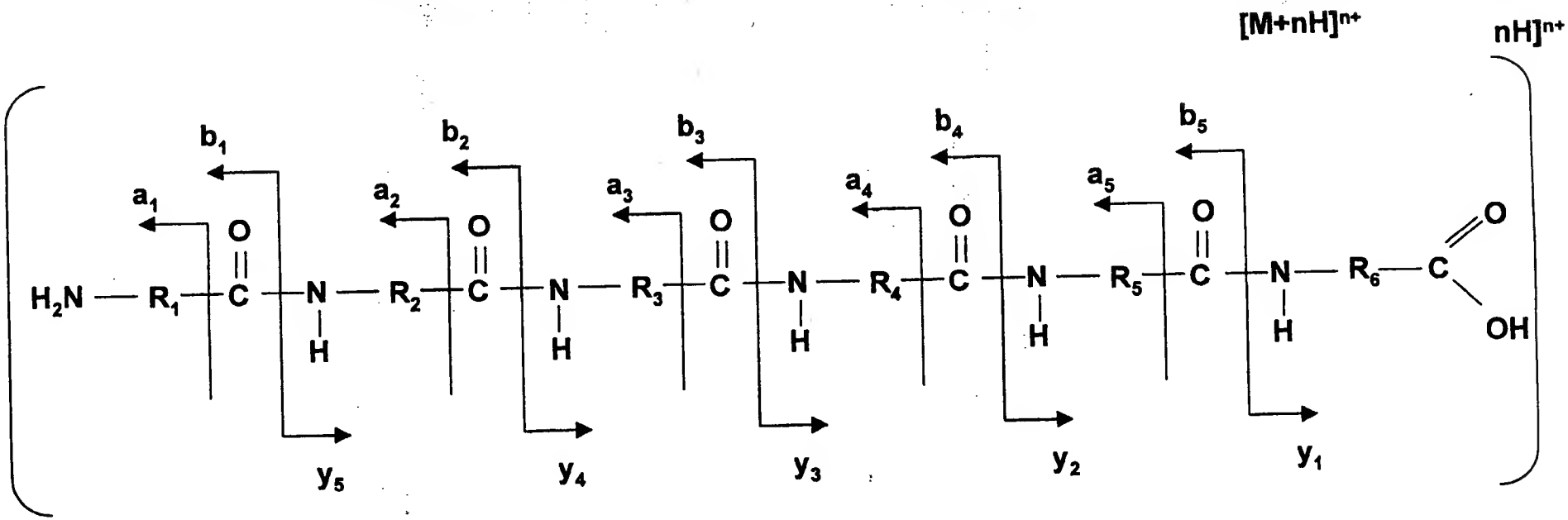


FIG 2. Method to Identify and Characterise Expressed Genes

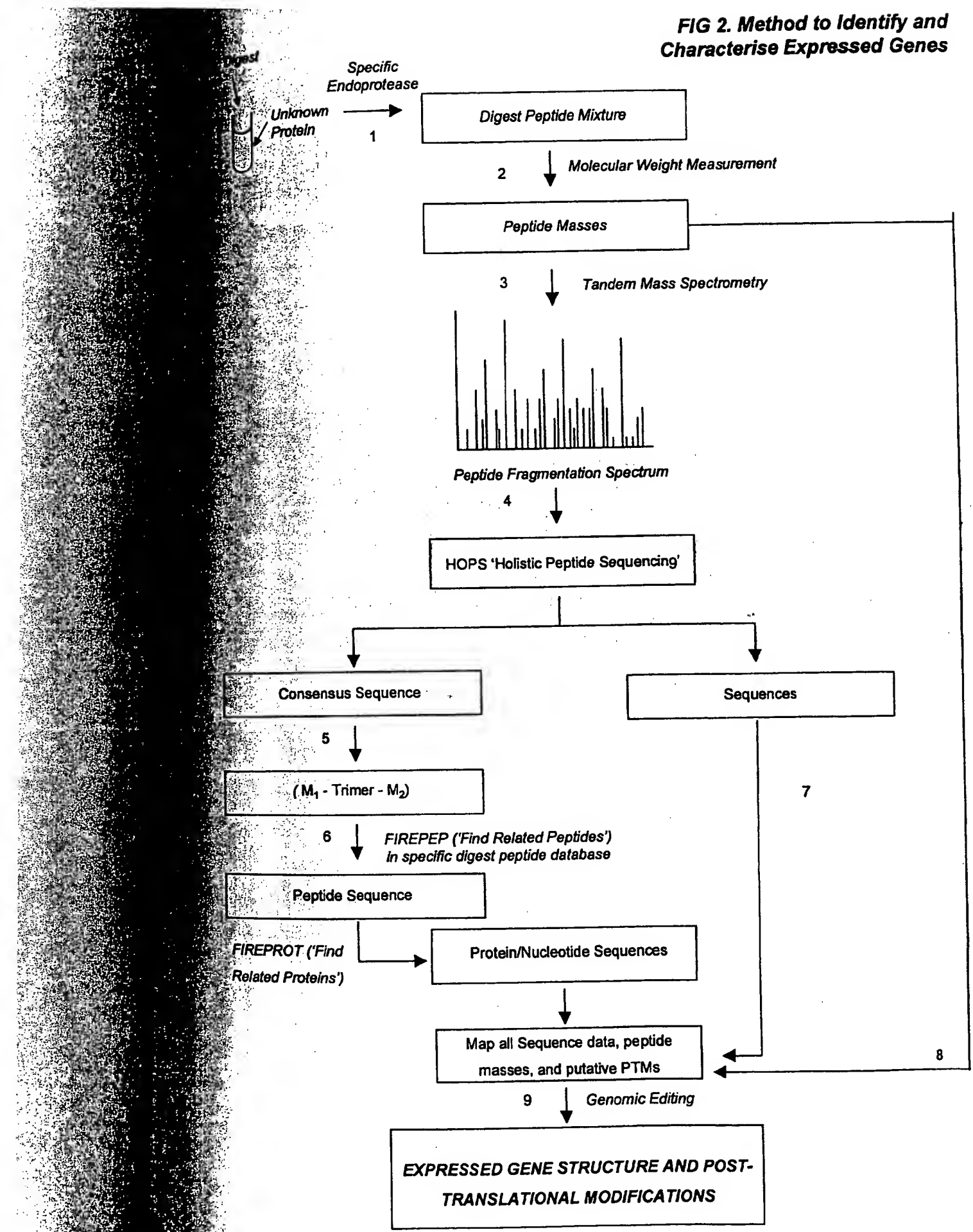


FIG 3. Overview of HOPS

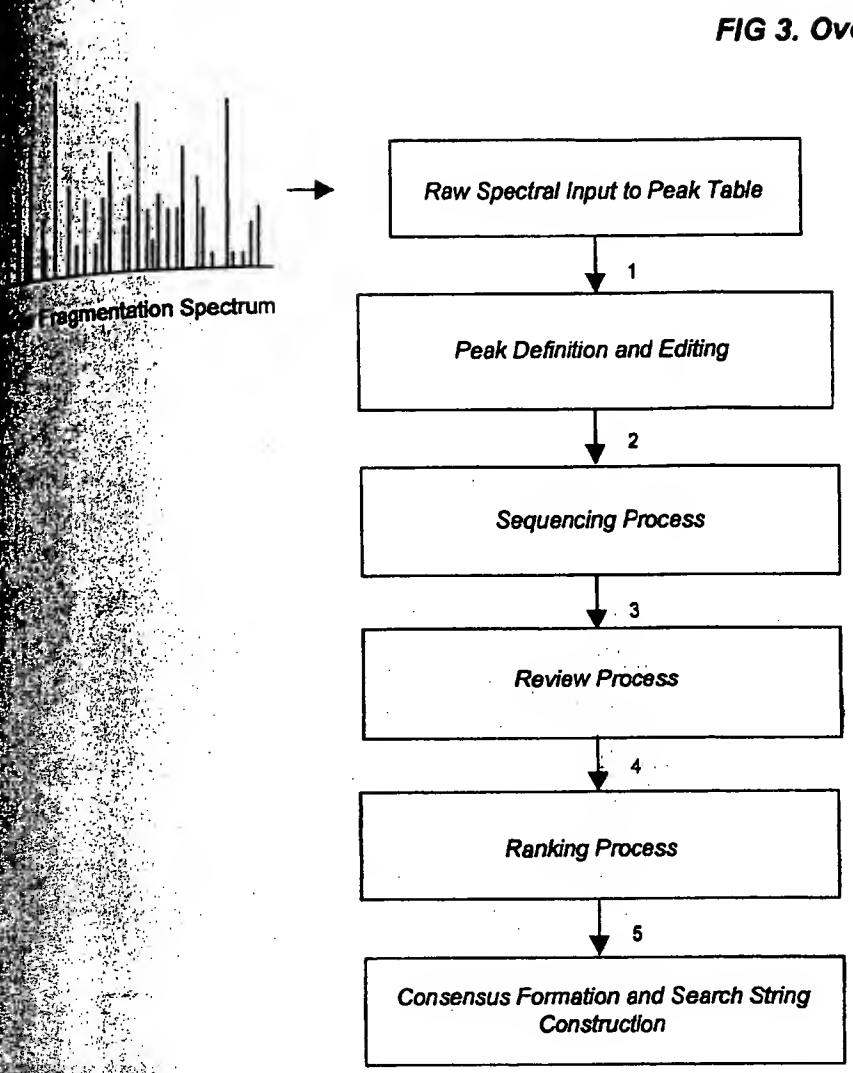


FIG 4. Peak Editing

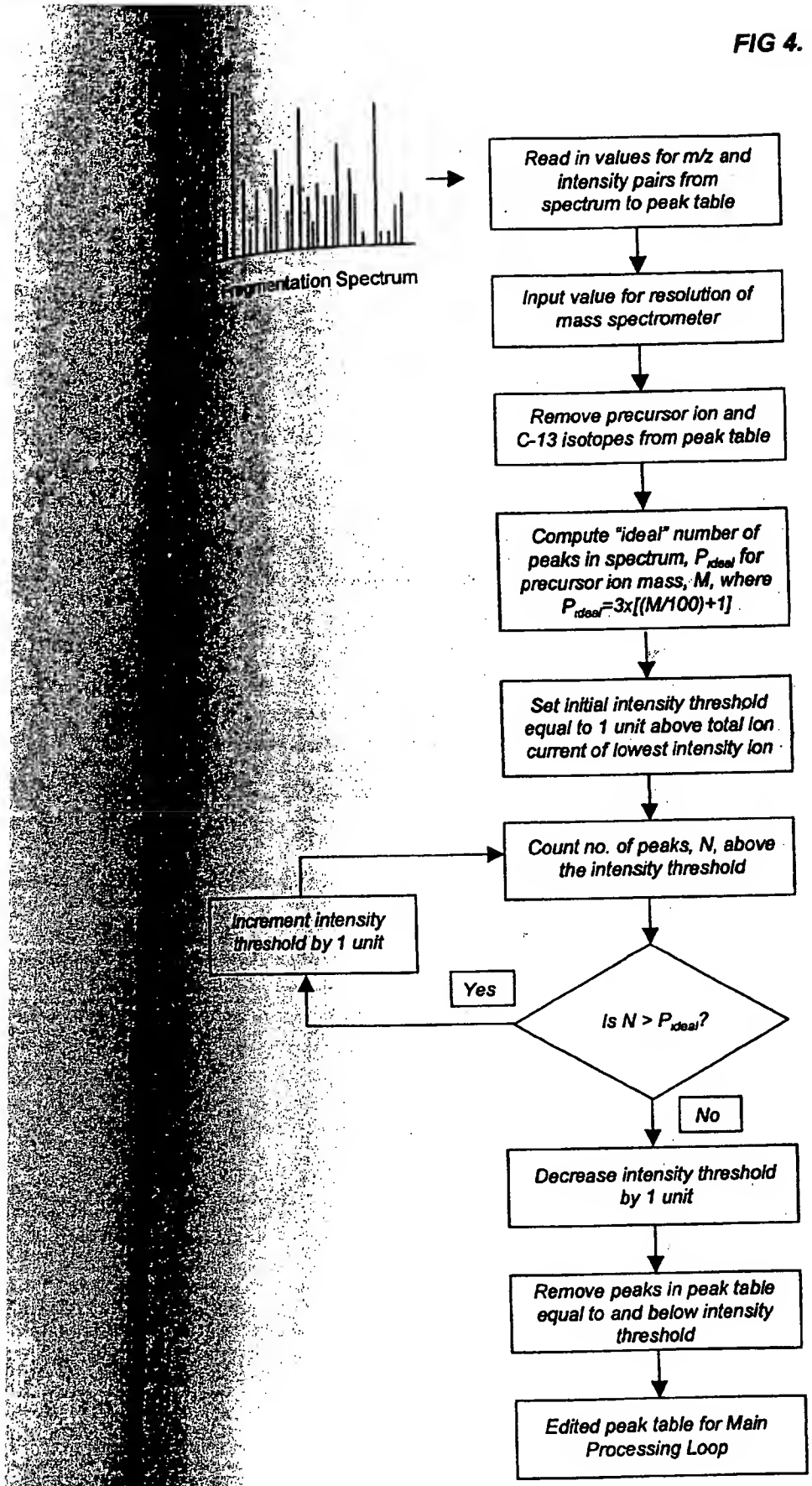


FIG 5. Main Sequencing Loop

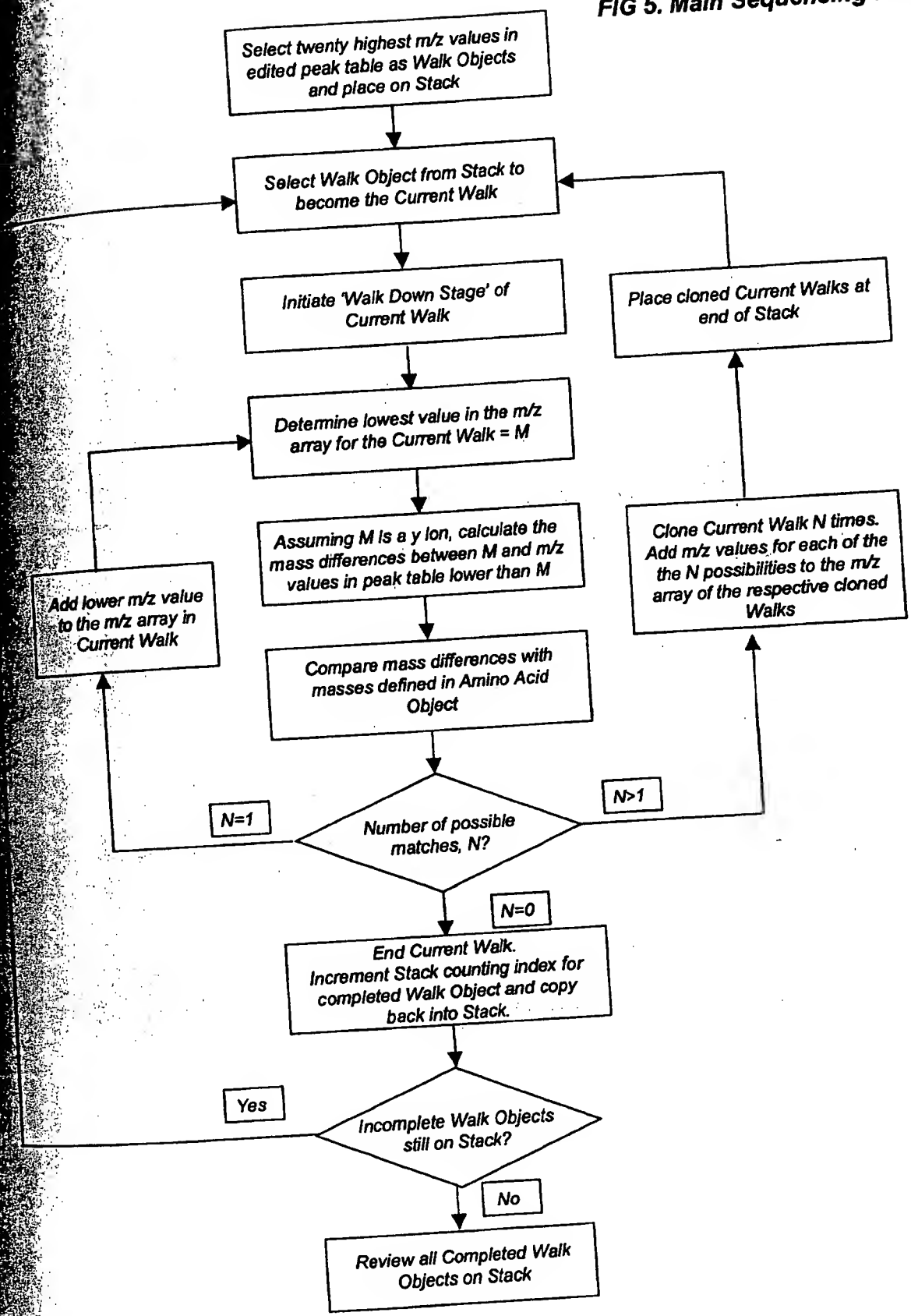


FIG 6. Review Process

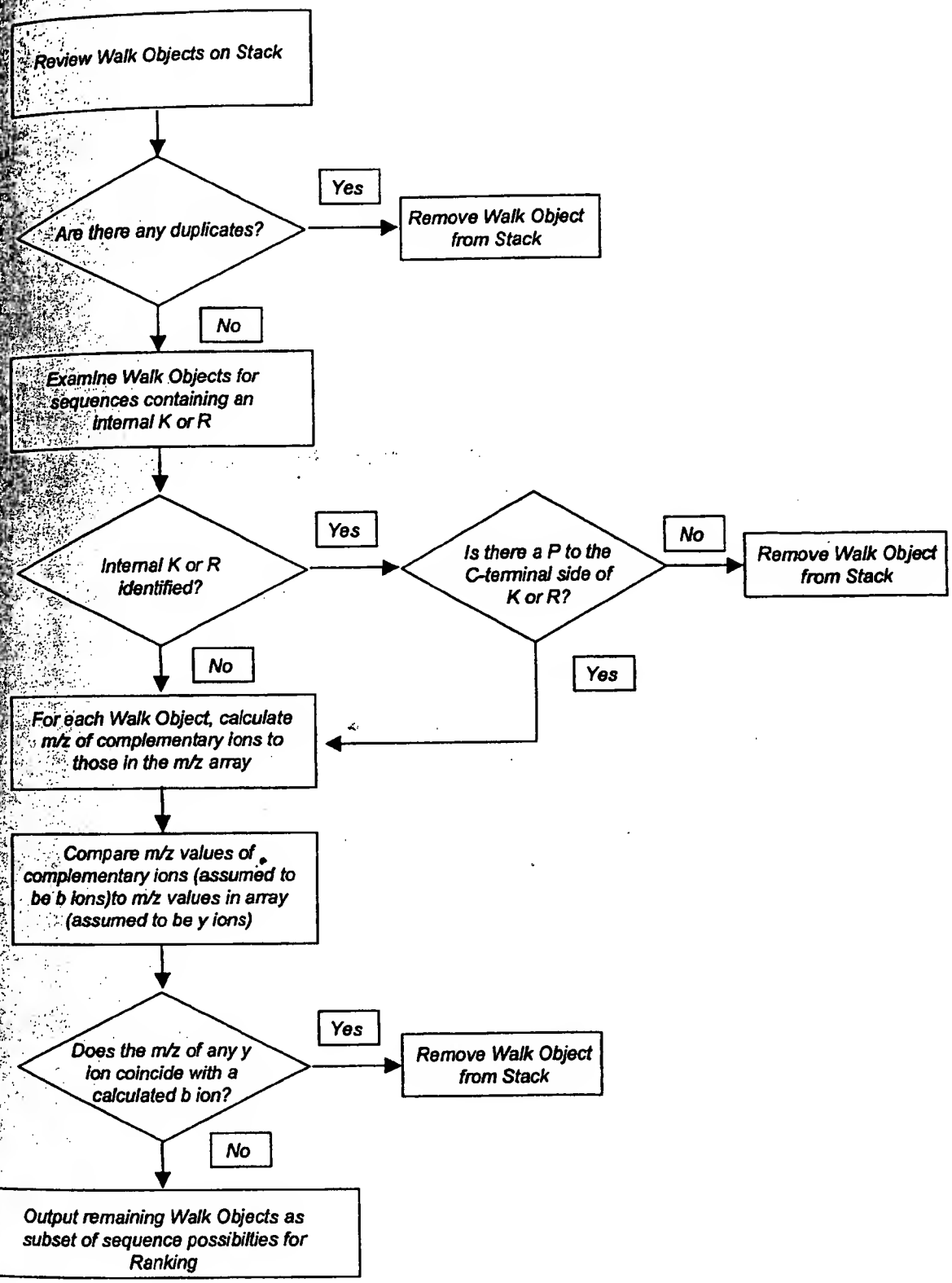


FIG 7. Ranking Procedure

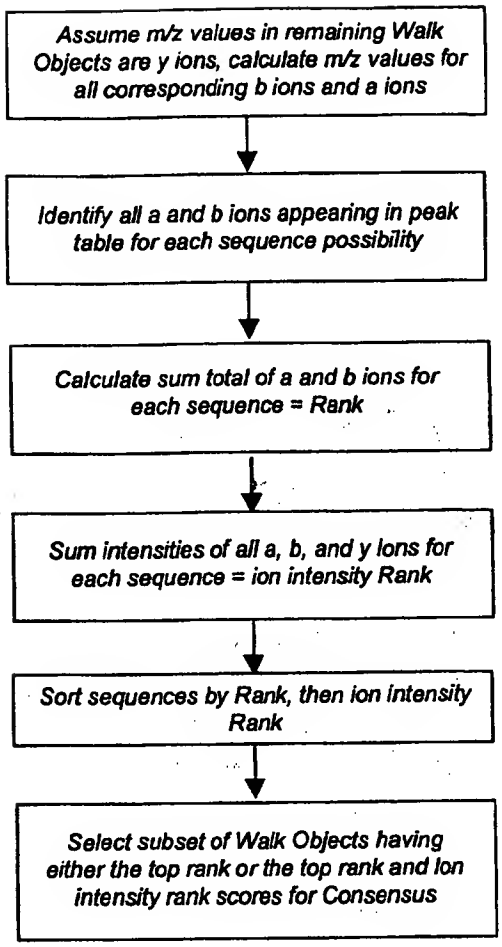


FIG 8. Consensus Procedure

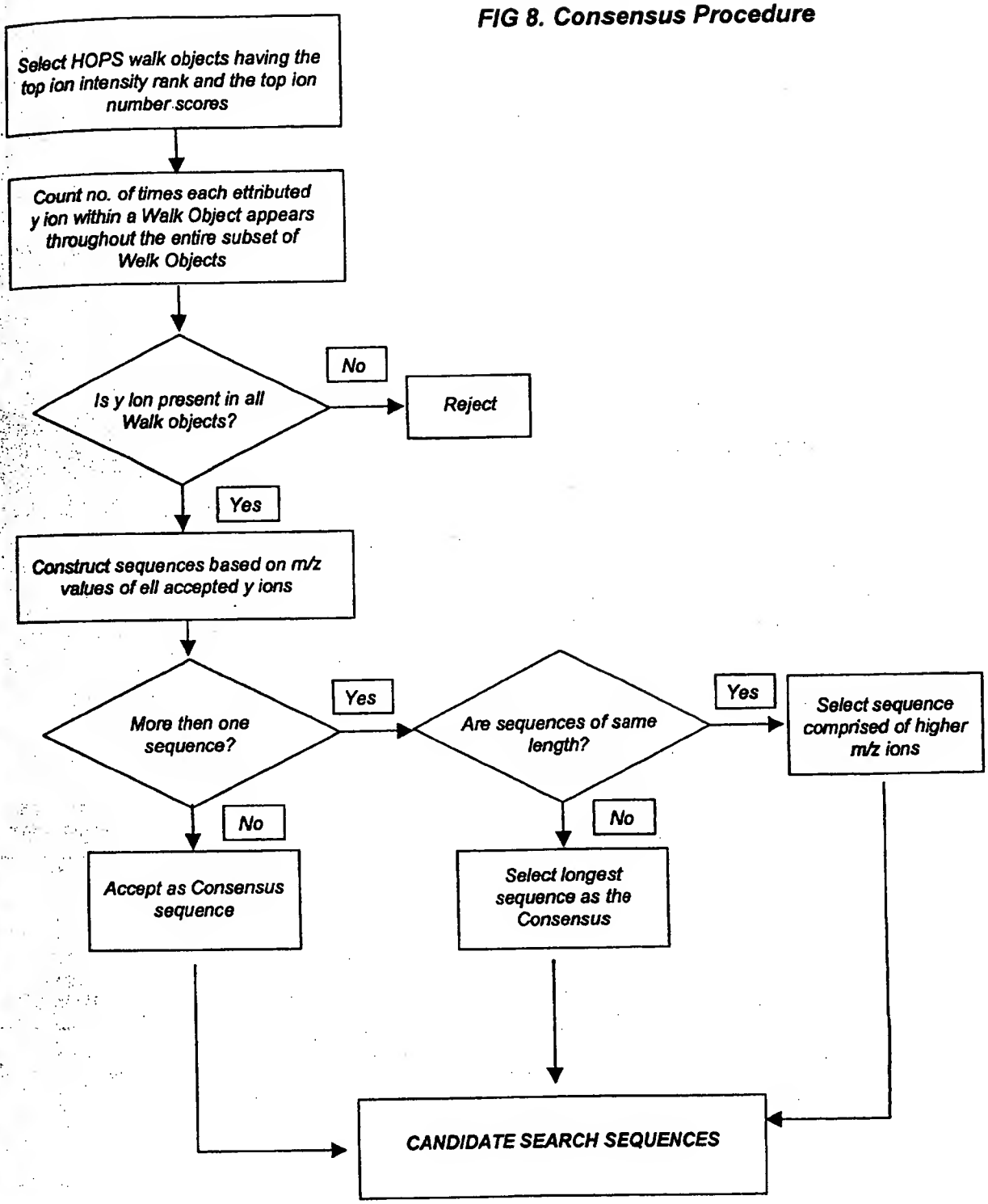


FIG 9. FIREPEP Method

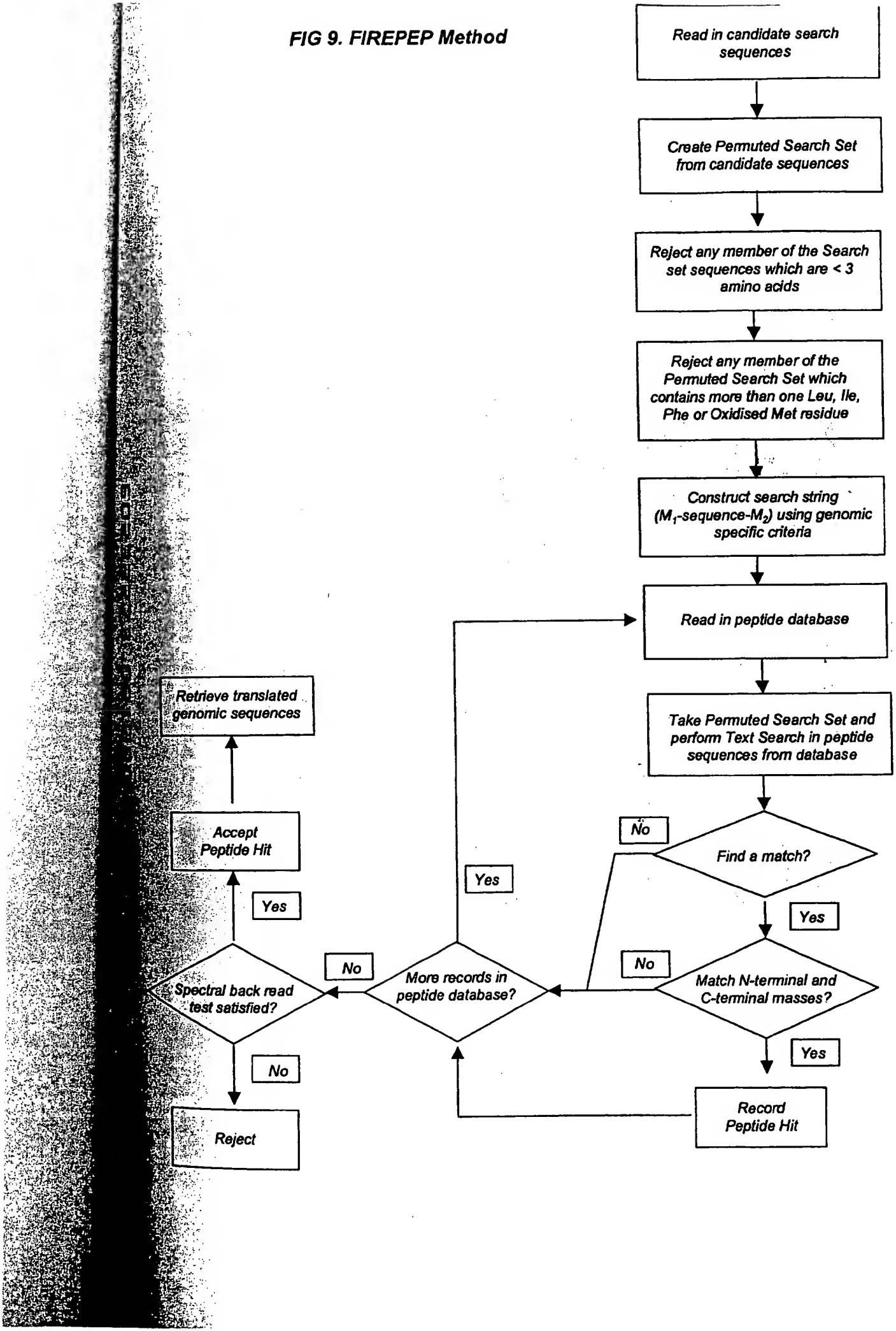
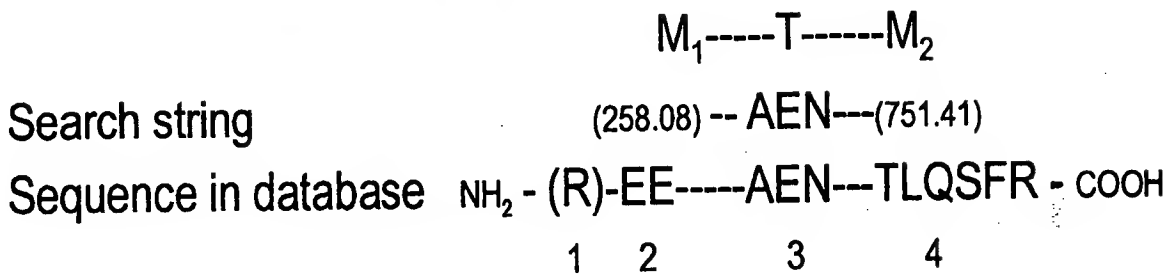


FIG. 10 Search string structure and database sequence attributes for unique identification of translated sequences from the human genome



Constraints for search string

1. M_1 - The value cannot equal the mass of a single naturally-occurring amino acid residue.

2. Trimer sequence- The three sequential amino acid residues taken from the HOPS candidate sequence. Only a single permuted residue (L/I, F/*M, or K/Q, if the K is followed by a P residue) is allowed within the trimer. The trimer sequence cannot contain only combinations of the residues V or A or a combination of either V or A residue and a single permuted residue. For example, IVA

3. M_2 - The mass of M_2 must be greater than 175.12.

Allowed database sequences

1. The nearest neighbor on the N-terminal of the retrieved sequence must be either a K or R residue.

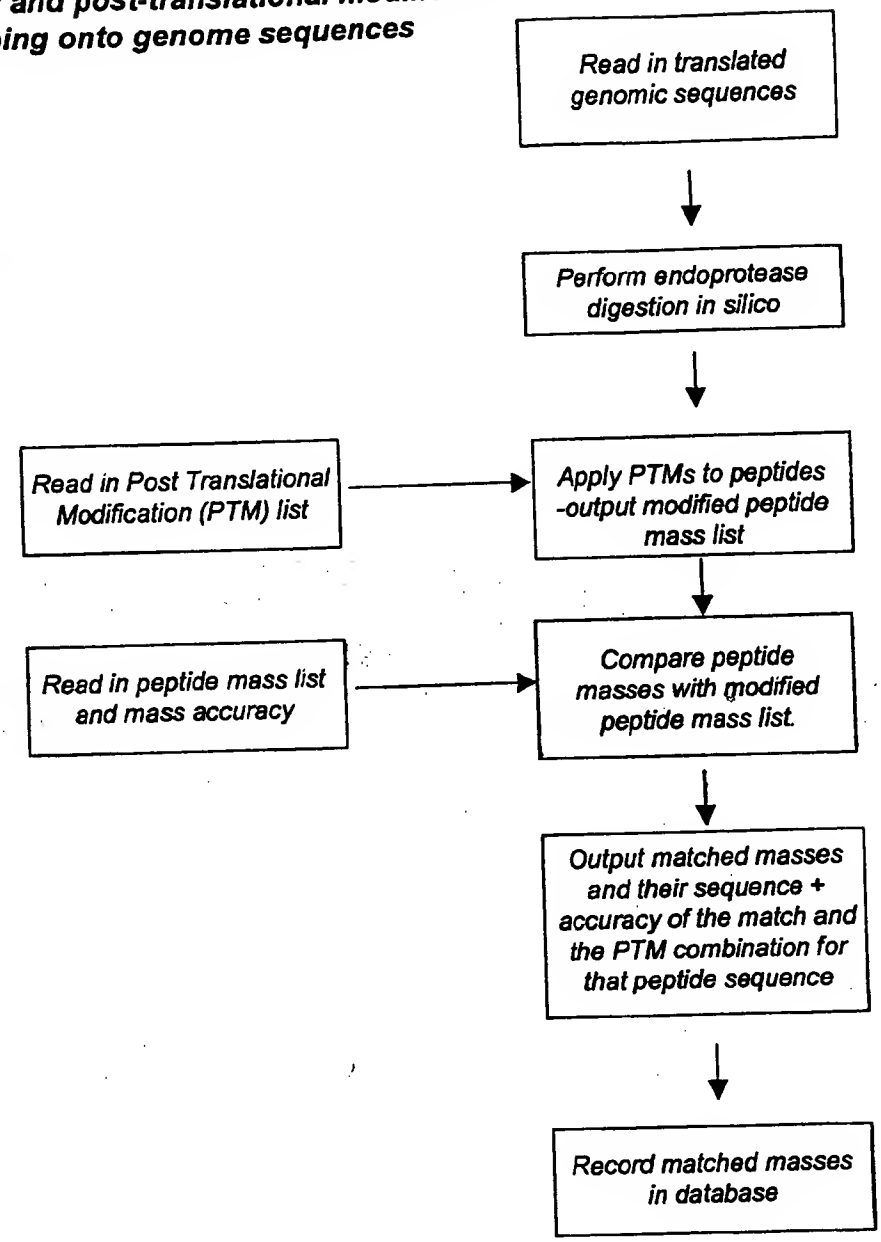
2. M1 cannot contain a K or R residue, unless followed by a P residue.

3. The peptide sequence must terminate in K or R and cannot contain additional K or R residues unless followed by a P residue (C-terminal).

Fig 12. Consolidation and Alignment of Genome Sequences using HOPS sequences

```
                                                
```

FIG 13. Mass and post-translational modification mapping onto genome sequences



PTM Flow Diagram

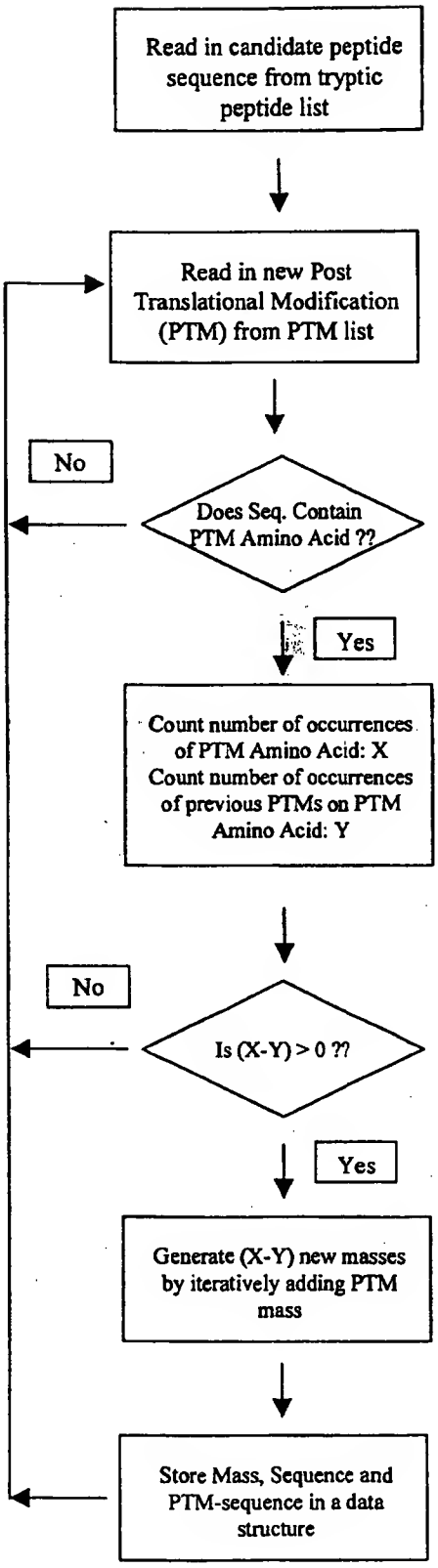
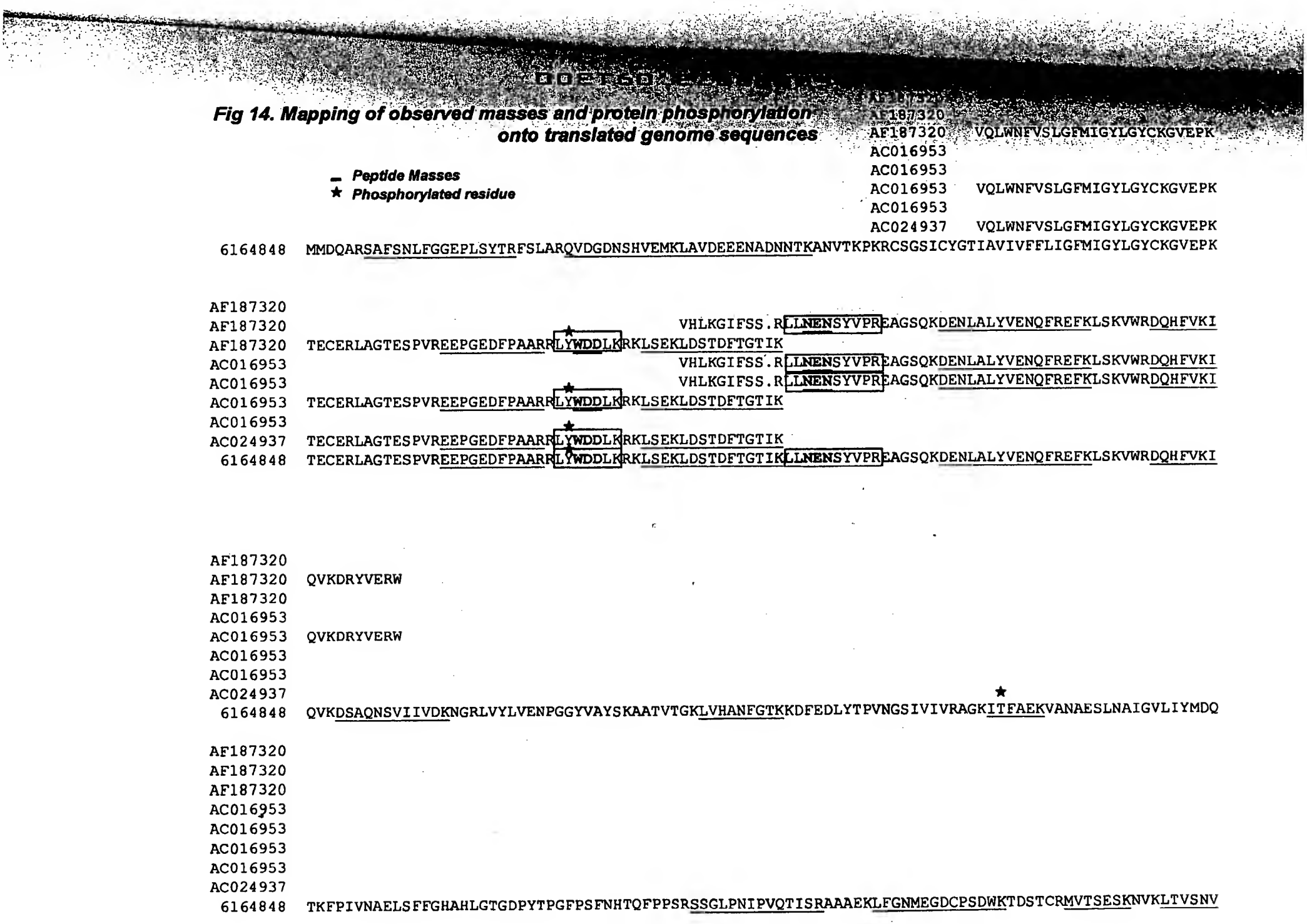


Fig 14. Mapping of observed masses and protein phosphorylation onto translated genome sequences



Accession	Frame	
AF187320	+1	
AF187320	+3	
AC016953	-2	
AC016953	-3	
AC024937	-1	
6164848		LKEIKILNIFGVIKGFVEPDHYVVVGAQRDAWGPGAAKSGVGTALLLKLAQMFSDMVLKDGFQPSRSIIFASWSAGDFGSVGATEWLEGYLSSLH

Accession	Frame	
AF187320	+1	PRIVSQDTDYPYLGTTM
AF187320	+3	
AC016953	-2	PRVVLQDTDYPYLGPTM
AC016953	-3	
AC024937	-1	
6164848		LKAFTYINLDKAVLGTSNFKVSASPLLYTLIEKTMQNVKHPVTGQFLYQDSNWASKVEKLTLDNAAFPFLAYSGIPAVSFCFCEDTDYPYLGTTM

Accession	Frame	
AF187320	+1	DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKVSTDSNYVFILLNVKYFEM
AF187320	+3	
AC016953	-2	DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKVSTDSNYVFILLNVKYFEM
AC016953	-3	
AC024937	-1	
6164848		DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKEMGLSLQWLYSARGDFFRATSRLTTDFGNAEK

Accession	Frame	
AF187320	+1	
AF187320	+3	
AC016953	-2	
AC016953	-3	
AC024937	-1	
6164848		TDRFVMKKLNDRVMRVEYHFLSPYVSPKESPFRHVFWGSGSHTLPALLENLKLRKQNNGAFNETLFRNQLALATWTIQGAANALSGDVWDIDNEF

```
AF187320
AF187320
AF187320
AC016953
AC016953
AC016953
AC016953
AC024937
 6164848  LKEIKILNIFGVIKGFVEPDHYVVVGAQRDAWGPGAAKSGVGTALLLKLAQMFSDMVLKDGFQPSRSIIFASWSAGDFGSVGATEWLEGYLSSLH

                                                                                                  PRIVSQDTDYPYLGTTM

AF187320
AF187320
AF187320
AC016953
AC016953
AC016953                                                                                          PRVVLQDTDYPYLGPTM
AC016953
AC024937
 6164848  LKAFTYINLDKAVLGTSNFKVSASPLLYTLIEKTMQNVKHPVTGQFLYQDSNWASKVEKLTLDNAAFPFLAYSGIPAVSFCFCEDTDYPYLGTTM

AF187320  DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKVSTDSNYVFILLNVKYFEM
AF187320
AF187320
AC016953                    TKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKVSTDSNYVFILLNVKYFEM
AC016953
AC016953
AC016953  DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKVSTDSNYVFILLNVKYFEM
AC024937
 6164848  DTYKELIERIPELNKVARAAAEVAGQFVIKLTHDVELNLDYERYNSQLLSFVRDLNQYRADIKEMGLSLQWLYSARGDFFRATSRLTTDFGNAEK

AF187320
AF187320
AF187320
AC016953
AC016953
AC016953
AC016953
AC024937
 6164848  TDRFVMKKLNDRVMRVEYHFLSPYVSPKESPFRHVFWGSGSHTLPALLENLKLRKQNNGAFNETLFRNQLALATWTIQGAANALSGDVWDIDNEF
```

INFORMATICS SYSTEM

A. Specific Aims

SurroMed's long-term research objective is to create a bioinformatics framework for the myriad of measurement technologies that will be used to create comprehensive differential phenotypes of humans using blood, other fluids or tissue. Differential phenotyping involves the creation of a "fingerprint" that can be used to differentiate patients into disease states or to monitor disease progression or the efficacy of drugs. To create a blood fingerprint, samples, collected over the course of disease from each patient, can be analyzed using cytometry and immunoassay techniques, and proteomic/ orgeomic (large and small molecule analysis) techniques including mass spectrometry. To allow differential phenotyping, the raw data collected must be preprocessed and interpreted, using a wide variety of measurement-technology dependent techniques and algorithms. Then, an integrated database is needed to warehouse all these data. Finally, a battery of data analysis/mining techniques is applied against this data. The results of this process include the discovery of biomarkers for predicting specific clinical endpoints and for assessing the efficacy of specific drugs. Our broad research goal is based on the following hypothesis — *measurements from multiple technologies can be combined into one database in such a manner as to allow robust differential screening in order to locate and identify unique novel biomarkers for disease.* This research plan specifically focuses on incorporating mass spectrometry data into our platform.

To enable use of mass spectrometry for differential phenotyping, multiple algorithms and tools will need to be developed and integrated into one cohesive system accessible by clinicians, biologists, chemists, and biostatisticians. The aims of this proposal are multifold. They include: i) development of conversion programs to combine mass spectral data from varied instruments, each having a unique file format and technological character; ii) development of a scalable database schema to support reduced mass spectra, experimental protocol, and sample tracking; iii) integration with the present architecture supporting cell cytometry data and immunoassay data; iv) development of algorithms to reduce large mass spectral data sets from different mass spectrometers which use different ionization techniques, as well as different separation techniques, in such a manner as to preserve enough information about rare chemical species to

allow for differential phenotyping of low concentration analytes; v) creation or adaptation of present clustering algorithms to segregate individual mass spectral peaks as potential biomarkers; and vi) design of visualization tools, both to steer chemical and biological assay development and for visualization of all multivariate data collected from SurroMed's diverse measurement technologies.

To determine the versatility and robustness of this integrated architecture, after the proposed tools are developed, we will put its full capabilities to test in the area of Type I diabetes mellitus. The immunological basis of insulin-dependent diabetes mellitus (IDDM) makes it an ideal disease to study using SurroMed's diverse repertoire of immunological tests and biological talent. Since the cellular arm of the immune system has been implicated in Type I diabetes, SurroMed's already developed technology for cellular phenotyping should be extremely valuable. Combine this with molecular phenotyping based on our Nanobar Code Identification Tags™ and mass spectrometry, the focus of this proposal, and one has a means to generate new clinical biomarkers for diabetes mellitus. The discovery and understanding of these biomarkers will lead to novel therapies for the disease.

B. Background and Significance

Recent advances in technology have made comprehensive patient phenotypic analysis an achievable objective in the near future. Such phenotyping will accelerate the pharmaceutical discovery and development process and enable the prevention, the precise diagnosis and the personalized treatment of disease. New technologies for comprehensive phenotyping – at the molecular, cellular and whole organism levels – are necessary to exploit recent advances in genomics, decipher gene function, identify root causes of disease and realize the promise of personalized medicine. Application of phenotyping to the discovery biological markers – novel characteristics of the patient samples that have a discrete relationship or correlation as an indicator of normal biologic processes, pathogenic processes or pharmacological responses to a therapeutic intervention – will be broadly enabling for a wide range of pharmaceutical and diagnostic discovery and development applications. In particular, utilization of biological markers in diabetics in the clinical trial setting to stratify and stage patient populations, and more

rapidly assess drug efficacy and safety, will result in significant improvements and efficiencies in the clinical drug development process.

Differential phenotyping – the comparison of bioanalytical and proteomic/orgeomic (protein and small molecule) data in samples from patients and controls across time – promises to be an important source of biomarkers, especially in diseases like Type I diabetes which are immunological in nature and are known to involve small metabolites. The advent of microfluidics, advanced laser-scanning cytometers like SurroScan™, Nanobar Code Identification Tags™, combinatorial separations, and a host of other new technologies enables the new differential phenotyping approach. These new technologies generate "increased (per volume) information content", a phrase which encompasses the reduction in the volume of sample required to carry out an assay, the highly parallel measurements ("multiplexing"), such as those involving immobilized molecular arrays and Nanobar Code tags, and the incorporation of multiple information channels, such as in liquid chromatography (LC)-electrospray (ESI) MS/MS. It is our belief that the last of these technologies, mass spectrometry, has not been fully utilized in differential phenotyping for biomarker discovery.

Differential phenotyping and mass spectrometry, in particular, lead directly to an increase in the amount and the complexity of information we can derive from patient samples. A one-hour LC/MS experiment can produce 30 MB of data. The character of this data and the orthogonal nature of the measurement technologies are necessitating the development of novel data handling paradigms. For example, the same sample run on different types of mass spectrometers generates different kinds of data that must be integrated to maximize the information content we can derive.

The informatics concepts involved in differential phenotyping represent a fundamental shift in the nature of data analysis and storage in proteomics. "Classical" proteomics involves assaying the quantities of a set of known proteins in blood using technologies like 2-D gels or mass spectrometry. The type of data produced by this type of proteomics is conceptually simple to handle and simple database schemas can be used to store it. In the mass spectrometry approach, for example, all spectral peaks not associated with the protein of interest can be discarded and only the quantity of known proteins is stored. Mass spectrometric software that has been integrated with a database follows this simple approach of storing only the quantities, derived from peak intensities, of the set of proteins that best match a set of previously measured

spectra along with the associated probability metrics. ProteoMetrics, MicroMass, Ciphergen and Genomic Solutions are examples of companies that have such software and databases. However, discarded peaks have significant information content and therefore valuable information is lost when following this "classical" approach. Differential phenotyping, while a more difficult approach with respect to data analysis and storage, offers much more abundant and rich information. In fact, there is a fundamental lack of software available to attack the problem of data handling, storing, and mining in differential phenotyping. Questions that need to be addressed, and that we intend to address, include: Which peaks are significant to store? Which multiple peaks represent a single protein? What is the best way to store a variable number of mass spectra in a database to optimize mining?

Genotyping and expression arrays are a second class of technologies that can be compared to differential phenotyping using mass spectrometry. However, current bioinformatic approaches to genotyping and expression arrays do not offer significant carry-over to differential proteomic/orgeomic phenotyping using mass spectrometry. Many genomics companies have highly integrated bioinformatic tools for analyzing and storing DNA sequence information, as do many public institutions. In the context of genotyping, these are typically sequences containing single nucleotide polymorphisms (SNPs) that can be compared across patients and diseases to elucidate genetic markers for disease. Sequenome, GeneTrace, and Affymetrix are examples of companies carrying out such research. While genetic codes are one type of biomarker, from an informatics point of view, they differ significantly in terms of retrieval, processing, storage and mining from proteomic mass spectral signatures. Genomic sequence database structures, like those at Genbank, are well defined and tools to access, to search for homology, and to process sequences in them, have been developed. The data they contain is one-dimensional – a string of nucleotides. In contrast, differential phenotyping and proteomic/orgeomic data, especially across orthogonal measurement systems, requires unique data-specific solutions and databases that have not yet been developed. The data can be noisy, highly multidimensional – the number of variables measured can surpass the number of samples by orders of magnitude – and vary in character – categorical versus continuous data for example.

Present commercial mass spectrometric software tools lack many of the needed features to do cross-platform assay development and differential protein and small molecule phenotyping. In order to perform broad-spectrum differential phenotyping of small, organic molecules and of

proteins, different mass spectrometers with different ionization methods are needed and, therefore, there is a need to integrate data from all the mass spectrometer platforms (ESI, electron ionization (EI), and matrix-assisted laser desorption ionization (MALDI)-based). Unfortunately, the software built around each instrument is currently inflexible and does not allow its application to data from another type or brand of mass spectrometer. There is not yet even a file format standard capable of handling the different types of mass spectrometric data and, in fact, the file formats are proprietary and trade secrets. As a result, even data extraction remains difficult. Additionally, each spectrometer supports different modes of operation that result in the output of data that differs significantly in character and dimensionality. Another missing feature is the ability to compare mass spectra and differentially analyze them. The analysis software is, instead, centered on protein identification. While a few programs exist that might satisfy the problem of analyzing difference spectra, they only analyze a single mass spectra snapshot and not the complete mass spectra collected over the complete retention time interval when doing LC/MS.

The lack of software that can perform mass-spectrometric differential phenotyping and store the volumes of complex data in a minable database, coupled with the problem of integrating data from other types of mass spectrometers and other measurement platforms (laser-scanning cytometers, enzyme-linked immunoabsorbant assays (ELISA), etc.) obviates the need for the development of an integrated set of software tools to speed analysis and contribute to the discovery of novel biomarkers. Only after this integrated software and database package is developed can we expect to achieve maximum utility from the recent technological advances in mass spectrometry and other technologies that can be used in conjunction with it, and can we expect to have better clinical diagnoses and treatments for diseases like diabetes.

C. Preliminary Studies

We will describe our preliminary experience and results in the following areas:

- Cytometry and immunoassay informatics platform;
- Mass Spectrometry data analysis;
- Nanobar Code Identification Tags™ separation analyzed with Mass spectrometry.

The first two sections explain previous informatics work carried out at SurroMed, while the last section explains the new experimental technology developed at SurroMed that has created the need for better mass spectrometry informatic tools.

Cytometry/Immunoassay informatics platforms

SurroMed has developed a significant bioinformatics infrastructure for the discovery of new biomarkers through application of novel data mining methods to a highly integrated database. The infrastructure currently in place has routinely been used to support clinical studies in many disease areas, including osteoarthritis (OA), asthma/allergy, and influenza, and can now be used for studying diabetes.

To date, SurroMed's bioinformatic accomplishments include a comprehensive suite of automated software for the design of measurement protocols, and the capture and interpretation of measurement data produced by a wide variety of measurement technologies. These measurement technologies include SurroScan™ (for cell expression analysis) and ELISA (for soluble factors analysis). Figure 1 diagrams the informatics platform that is in place for our SurroScan™ technology.

The scalable database architecture developed at SurroMed integrates a wealth of information spanning many disciplines that range from clinical science to biology to chemistry. At the same time, the database is structured in such a way as to retain sufficient flexibility to accommodate future measurement technologies. This integrated database is a key enabler to rapid and powerful data analyses. We have developed an integrated suite of tools for efficiently querying, summarizing, and performing advanced statistical analyses on the integrated data. For example, we have developed methods for identifying new patterns in laboratory data that have strong links to given disease states, as well as detecting statistically significant differences between biosample populations. Other software tools support hypothesis exploration and

validation for such areas as testing drug efficacy, and mining massive bodies of measurement data for reliable markers when only a small sample population is available.

Integrated database

SurroMed's scaleable database architecture allows the capture and integration of both clinical and laboratory data. This architecture is perfectly suited to handle the large volumes of data generated via longitudinal, comprehensive phenotyping of small numbers of patients, because it was designed with that goal in mind. The ability to handle cellular data, immunoassay data, and clinical data now in the same database and mass spectrometric and genomic data in the future, as well as to carry out sophisticated data mining across the entire database, are unique and powerful features that are unavailable in any other database of which we are aware.

With flexibility in mind, our database is organized in a highly modular fashion. Each module contains a set of tables that implement a particular information model. The

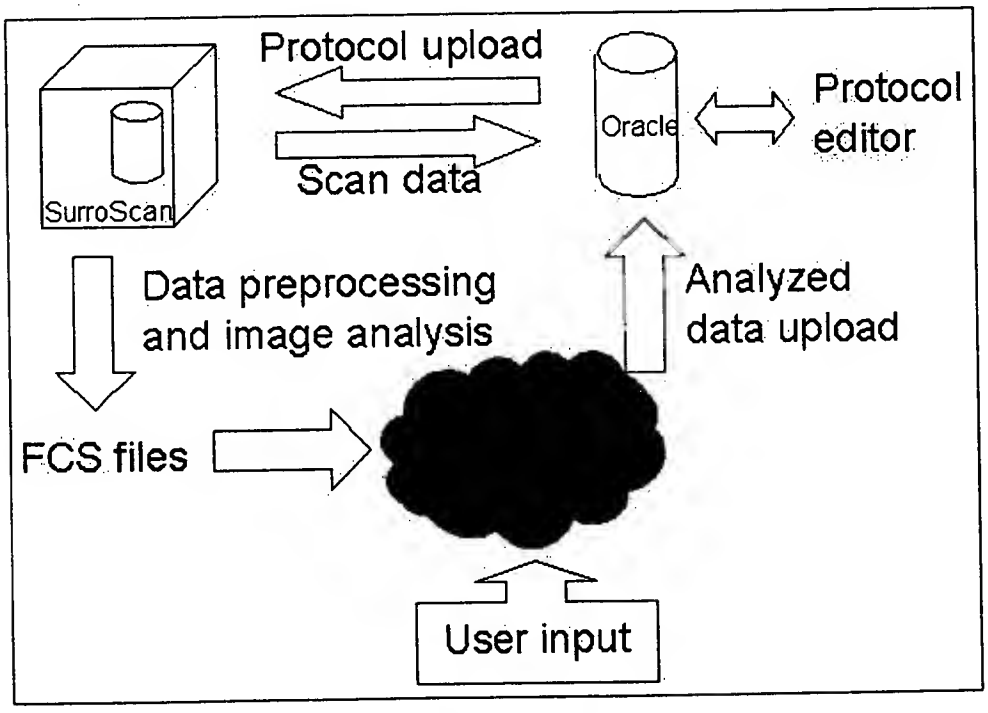


Figure 1: The SurroScan™ cytometry informatics platform.

information models represent self-contained units of information and are designed so that their interdependencies are minimized. Some of these information models are not specific to any

bioanalytical discipline. For example, the Clinical Study Protocol model captures the design of a clinical study that includes annotations on biosamples that are useful to consider for the purpose of the study. Other information models are specific to the bioanalytical disciplines and their contents are idiosyncratic to each discipline. For example, we have 3 submodels for cytometry, 3 submodels for immunoassay, and 2 submodels for clinical information. This approach to data integration makes it easy to quickly adapt to changes in existing measurement technologies as well as adopt new measurement technologies such as mass spectrometry and genomics.

Protocol development and storage

When carrying out a clinical study, it is vital that information about the protocols used to collect data be archived, preferably with the data itself. We have developed informatics tools that achieve this aim. These tools allow experimentalists in the laboratory to easily enter new or create new versions of existing protocols in our central database. When data is collected, a protocol is chosen and our tools store pointers to the protocol in our database, thus linking the data to the protocol seamlessly.

For the cytometry platform, the protocol editor is used to define the clinical study protocol that is run on the SurroScan™ Instrument. The information can be thought of as 3 separate submodels: the AssayProtocol submodel, the ProtocolCartridge submodel and the PatientSample submodel. The AssayProtocol submodel is generally defined first, since it corresponds to tested and verified cellular assays. This submodel defines all the information needed for the data collection (sample preparation and scanning), data preprocessing (image analysis), and data format conversion. The ProtocolCartridge submodel defines what capillaries hold the different biosamples and uses the specified AssayProtocol. The PatientSample model is used to track the origin of the biosamples used in the clinical study and to associate that information with clinical data. These defined protocols are then loaded from Oracle into the local database on the cytometer being used to run the experiment.

For the immunoassay platform, we have written software, SurroMap™, that allows our scientists to develop protocols directly in a 96-well plate format, where plate maps are defined to hold both control samples and clinical samples in each virtual well location. Each experiment is then linked to a plate map, allowing results to be tagged with the correct biosamples.

SurroMap™ also allows scientists to define the various reagent and sample dilution factors used in the immunoassay protocol.

Data collection

After a protocol is chosen and laboratory scientists run an experiment, raw results are both stored in our database and saved as files for processing. We are currently in the process of automating the raw data storage, so that at the end of an experiment, after a brief look at the data for validation, the data automatically is loaded into Oracle. As a first step, for example, we have written a program to quickly display the raw cytometry data directly to the computer controlling the instrument. This is challenging in that we need to display 32 sets of data (one for each capillary), each of which can contain approximately 10,000 data points (cells), very quickly – before the experimentalist moves on to the next experiment. This allows the experimentalist to immediately check whether an experiment produced useful data, before discarding the sample and proceeding with other experiments. After this validation step, the experimental parameters are automatically uploaded into Oracle.

Data interpretation

Data interpretation is decoupled from data collection in order to allow scientists to analyze results when they have the time, rather than requiring the analysis immediately – a scalable solution to this issue. Over time, all data interpretation will be automated and only unusual data sets will be flagged for manual checking. Now, the process has both automated and manual steps.

Routines have been written to export data from the cytometers into a format that is readable by FlowJo™ (TreeStar Inc., CA), the commercial package we use to analyze our cytometry data. In FlowJo™, gates are chosen manually to classify cell events into cell-type specific populations. Once these gates are chosen, we have software that can automatically apply them to all samples for that assay, and the results are automatically loaded into Oracle. Current work focuses on automating the gate-calling procedure to use data from all experiments run and then automatically apply those gates to all samples simultaneously and store all data into Oracle at once. This work will eliminate the need for FlowJo and for user input, except in unusual cases, and is being developed using both unsupervised learning techniques and

knowledge-based techniques. These techniques will be evaluated and compared in terms of efficiency and accuracy of data interpretation.

Immunoassay results are currently analyzed using curve-fitting procedures that we have implemented. For example, users can choose to analyze controls with cubic spline, Chapman 4, or Sigmoid 5 curve-fitting algorithms, and can choose which control points to use or leave out in analysis. Once the controls have been analyzed, all assays on the same plate are analyzed with the parameters derived, and results are automatically stored into our central database.

Data analysis and mining

Once processed data is stored into our database, the data must be turned into information through data mining. Our database design was developed with the idea of performing data analysis and mining in mind. Several in-house experts, both biostatisticians and database experts, have developed and/or implemented statistical and other data mining algorithms with web-based user interfaces and either SAS, statistical libraries or other data mining engines as analysis tools that do the number crunching. For example, on the statistical front, we have tools to search for biomarkers by doing simple statistics, Fisher Discriminant (1) analyses, whose user interface is shown in Figure 2, and step-down Bonferoni (2) analyses. These tools are all web-based and allow clinicians, biologists or chemists within the company, or collaborators working with us, to access the data and analyze it. We are currently implementing a support-vector machine approach and other machine learning techniques to partition our patient populations and search for biomarkers. Lastly, we are implementing techniques for mining broad data sets for good biomarkers, a problem commonly faced by biotech researchers that involves very high combinatorics. Recognizing that a primary source of SurroMed's value is the ability to turn data into information, we are devoting a large effort to our database technology and data mining.

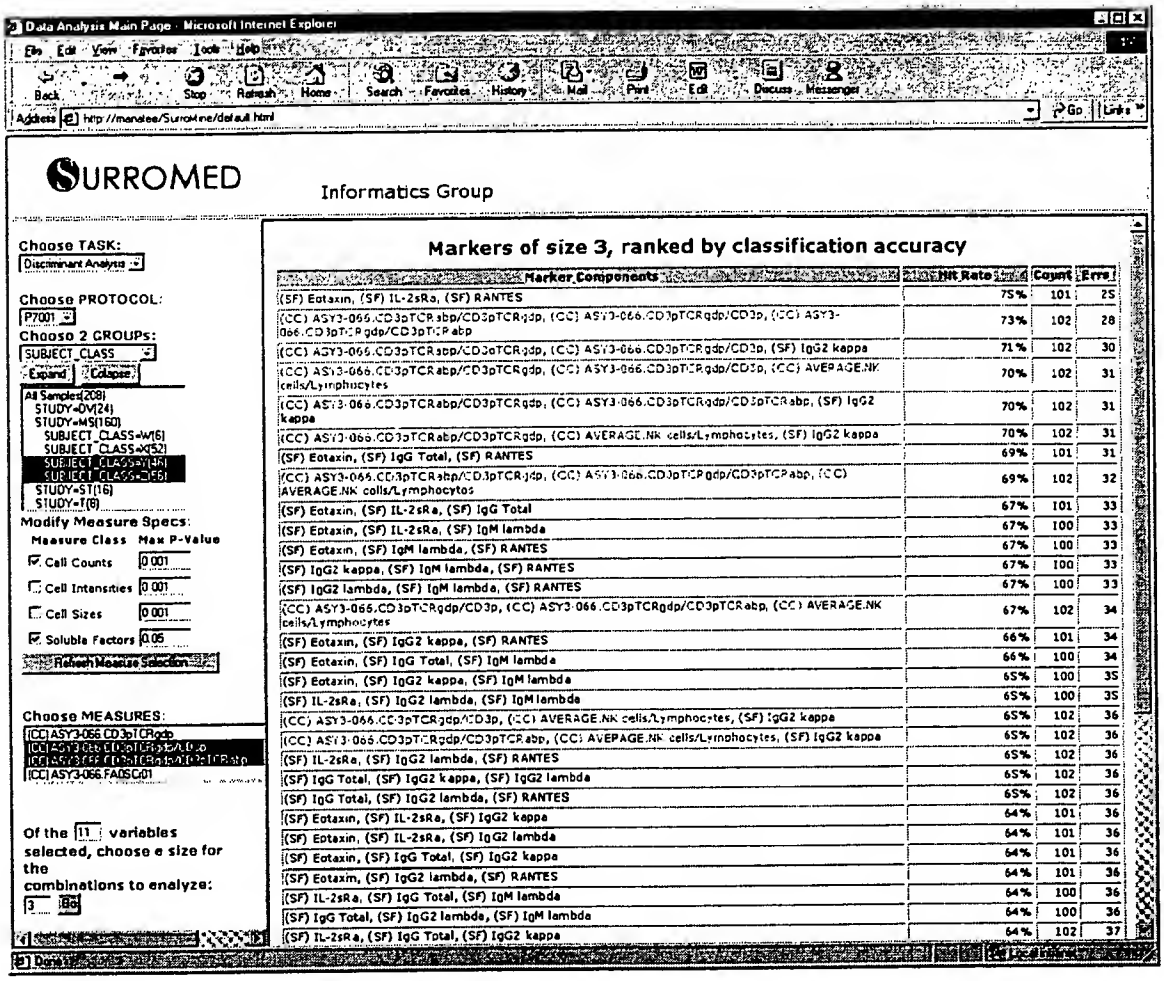


Figure 2: The interface of our Discriminant analysis tool used for mining broad bioanalysis data sets for biomarkers.

Mass Spectrometry data analysis informatics

Mass spectrometry is the newest addition to SurroMed's toolbox for the discovery of novel biomarkers through differential protein and organic molecule phenotyping. Our rationale for choosing mass spectrometry as a critical detection element for differential molecular phenotyping is predicated on its applicability to gather information for both high and low molecular weight species. Indeed, it is the only technique capable of both furnishing molecular identification of peptide fragments associated with large proteins and molecular weight/identification of molecules in the 100-500 atomic mass unit (amu) range. Moreover, insofar as several masses can be identified simultaneously, mass spectrometry is an intrinsically multiplexed detection technique. The multiplexed nature of mass spectrometry makes it a rich source of information, yet presents many technical challenges in terms of establishing a

bioinformatics infrastructure similar to that in place for cytometry and immunoassays. We have begun preliminary work in the areas of collecting and processing mass spectral data suitable for storage and mining in our central Oracle database.

SurroMed has acquired three mass spectrometer instruments with plans to purchase a fourth instrument in the near future. These consist of: a Finnigan LCQ-DECA ion trap-based electrospray (ESI) instrument; a Kodiak 1200 triple quad electron ionization (EI) and chemical ionization (CI) mass spectrometer from Bear Instruments; and a Voyager DE-PRO matrix-assisted laser desorption/ionization (MALDI) instrument from PerSeptive Biosystems.

Mass spectrometry file formats from different instruments are not standardized or publicized, as described in *Section B*. Our initial endeavor was to find a file format that could support all the modes of operations from each manufacturer's instrument. Bear Instruments' mass spectral file format was found to be flexible enough to support data produced by various modes of operation from the other two instruments, and we have partnered with them to support assay development and to provide some software development resources. Through our partnership with Bear and using component object modules (COMs) provided by Finnigan, we have developed C++ software to convert our Finnigan and PerSeptive data formats to the Bear format.

A typical mass spectrometer run of one hour in length on the ESI instrument can generate extremely large data files (typically 15MB to 80MB). From both a storage and mining perspective, this data needs to be reduced in such a manner as to retain its information content, yet discard noise. We have begun investigating algorithms for noise reduction as well as peak extraction methods including Biller-Biemann (BB) and extensions (3-4), the Windowed Mass Selection Method (5), Singular Value Decomposition (SVD) (6), the Component Detection Algorithm (CODA) (7), the Sequential-Paired Covariance (SPC) (8), Higher Order Sequential Paired Covariance (HO-SPC) (9), Backfolding (10), and Principal Component Analysis (PCA) (11). Using Mathwork's MATLAB, a high-level numerical language, these algorithms have been implemented and tested. These traditional methods for processing LC-MS data are designed for small molecule mixtures in which the level of analytes is relatively high, the noise being dominated by contributions from the LC mobile phase. It is desirable to identify individual components on the basis of their mass spectra including ions arising from fragmentation of their parent ion. These methods improve the appearance of the total ion current (TIC) plot by

removing mass chromatograms containing only noise and mass chromatograms with high background (i.e., CODA) or by sharpening individual component peaks (i.e., Backfolding). Component mass spectra are then identified from TIC peaks by calculating precise elution times for each ion peak (4) and finding components that co-elute (i.e., Biller-Biemann). In all of these cases, we propose that resultant pre-processed component mass spectra be stored in a database for future mining.

A problem with the above methods is the fact that they assume the noise in a spectrum is normally distributed. Our experience has shown that, in fact, the noise spectrum is non-normally distributed and therefore, we have developed a windowed-median-filter algorithm (a nonlinear filter) to remove noise. Initially results are quite promising (Figure 3) and we plan on continuing to explore this issue.

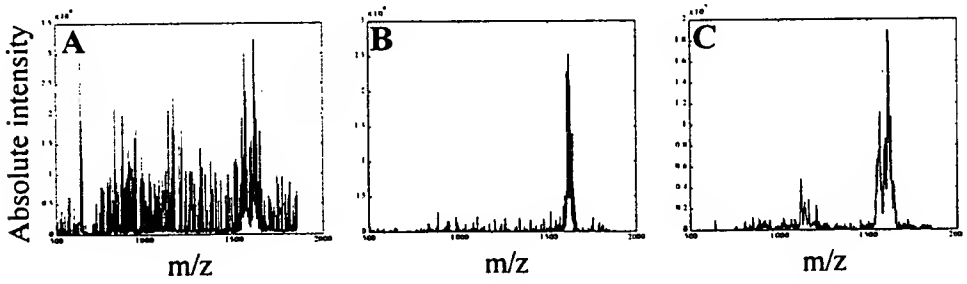


Figure 3: Preliminary results on preprocessing electrospray ionization (ESI) mass specrtral data. (A) represents the original total ion current (TIC). (B) represents CODA-(Similarity index is 0.8, averaging window is 5 points) processed data and (C) demonstrates that application of a custom median filter before CODA processing brings out suppressed components near m/z = 1150.

One key component of differential phenotyping using mass spectral data will be the algorithm chosen to recognize differences in mass spectra between biosamples drawn from different patients. Clustering algorithms, including principal component analysis (PCA) (12) and neural networks (13) (optimal for nonlinear models) hold promise. We have purchased a software package, HighChem Mass Frontier™, that supports principal component analysis of single mass spectra. While our needs will require a package that can analyze a complete mass spectrogram, this package offers some insight into the capabilities of this algorithm.

Nanobar Code Identification Tag separations analyzed with mass spectrometry

In the sections above, we have described informatics components of SurroMed's cytometry and immunoassay platform as well as some preliminary studies we've carried out for mass spectral analysis in differential phenotyping. Advances in Nanobar Code technology (14), combinatorial separation (combisep), and solid phase microextraction (SPME) (15) at SurroMed have made it possible for us to add a Nanobar Code/mass spectrometry component to our differential phenotyping platform. This section explains those advances.

State-of-the-art technology in proteomics relies on 2-D gel electrophoresis that allows separation of complex protein mixtures expressed at the level of whole cells, tissues or whole organisms. After 2-D gel electrophoresis and gel staining, the revealed protein spots are excised, extracted from the gel and subjected to enzymatic digestion. The resulting peptide fragments are then characterized by mass spectrometry (MALDI-TOF MS or ESI-MS).

At SurroMed, technology has been developed to carry out a similar procedure, but in a much-improved manner. The 2-D gel separation step has been replaced by separation methods that take advantage of our Nanobar Code Identification Tag platform and we have coupled those novel separation methods to both MALDI- and ESI-MS for detection of both small molecules and proteins. This technique is a vast improvement over current state-of-the-art proteomics for several reasons. Progress in the field has been hampered by the lack of reproducibility of the 2-D gel process, difficulties in protein quantification, and complexities associated with sample extraction. 2-D gels also suffer from a separation bias against proteins of very low and very high molecular weight. Accordingly, 2-D gels are incapable of profiling small organic molecules, as well as very large and membrane bound proteins.

Nanobar Code Identification Tags are cylindrical metal nanoparticles in which the composition along the particle length can be varied in a stripe-like fashion. Because the number of stripes, the width of stripes, the identity of the metals, and the overall particle shape can be varied, trillions of unique "flavors" can theoretically be produced. At SurroMed enough flavors have already been produced to carry out the work we envision and describe here. An electrochemical-deposition based synthesizer engineered and constructed at SurroMed is currently capable of making up to a hundred flavors, and we are investigating alternative techniques for making hundreds more flavors.

Analogous to conventional barcodes, which are read using the differential contrast of black and white stripes, Nanobar Code tags can be identified using conventional optical

microscopy, based on the pattern of differential reflectivity of adjacent metal stripes. Additionally, tags can be read using fluorescence microscopy, once they are derivatized and bound to fluorescently tagged biomolecules. This allows quantification of the bound biomolecule. Additionally, different stripes have been successfully derivatized in differential manners, allowing fluorescent images to identify the Nanobar Code tags as well. A first generation Nanobar Code Identification Tag reader has been engineered and built at SurroMed. First-generation image analysis software, shown in Figure 4, has been developed to locate and identify the Nanobar Codes.

Importantly, Nanobar Codes tags can be functionalized with proteins, nucleic acids, or small molecules — in short, any chemistry that can be carried out on beads, nanoparticles, surfaces or any other solid phase can be replicated on Nanobar Code tags. Combinatorially designed surfaces to capture several analytes simultaneously on distinctly coded particles from biological samples have no precedence in the separation sciences or the clinical chemistry arena. To generate these surfaces, we have used self-assembled monolayers (SAMs) terminated with reactive functional groups that have been derivatized with libraries of reagents to give Nanobar Codes an extraordinary variety in surface chemistry. The strategy of selectively binding to a functionalized surface a species from a mixture leads to their enrichment on the surface, generating a route to detection with higher sensitivity. These combinatorially-derivatized nanoparticles present surfaces with varying avidity for binding to the wide variety of molecules present in a biological sample and therefore are an excellent method for sample separation. The affinity capture techniques with these nanoparticles, unlike the chip-based systems, will use off-line incubation steps for capturing the analytes. Such an approach is not only inherently superior from a kinetic viewpoint (fast capture of analytes), but also advantageous from mass action laws to drive binding as the density of the binding determinants can be varied to accommodate a wide range of analyte concentrations encountered in a biological fluid. An approach like this can generate surfaces of the same magnitude as the number of molecules present in a biological sample. This is essentially one of the best possible modes of sample preparation prior to analysis.

Derivatized Nanobar Code identification tags can be used for complex sample separation, and then coupled to mass spectrometry to identify and quantify all components of all fractions of the original sample. Surface enriched laser desorption ionization (SELDI) (16) commercialized

by Ciphergen, is a variation of MALDI based on the principle of using functionalized surfaces as routes to enrichment and detection. Ciphergen offers 5-6 different surfaces upon which protein

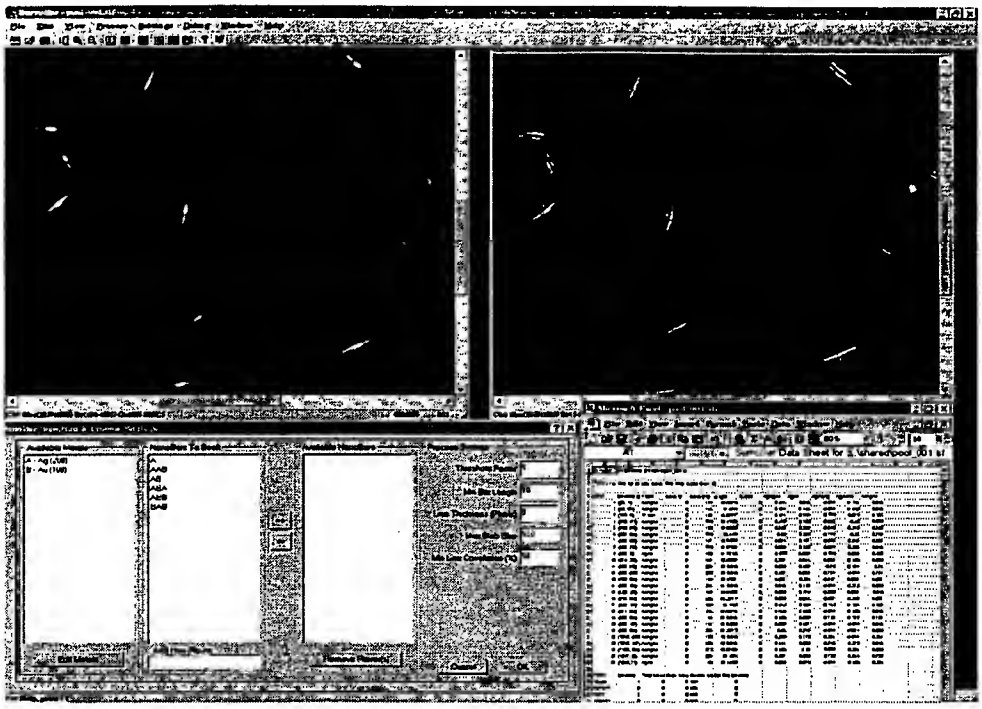
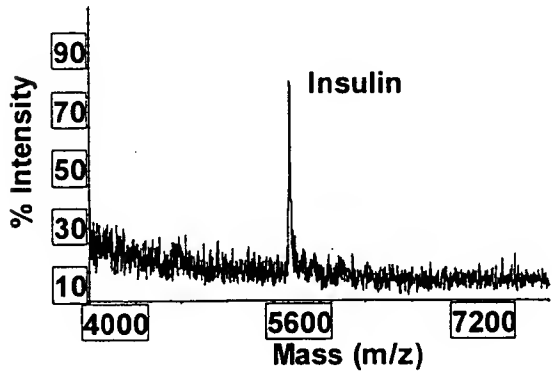


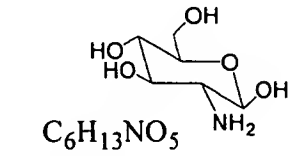
Figure 4: A screen shot of the interface to our Nanobar Code tag image analysis software. The top left frame shows the raw optical image; the top right frame has the analyzed image with numbered tags; the bottom left frame allows the user to describe the types of Nanobar Codes to be identified and the bottom right frame shows assignments for each identified tag and associated probabilities.

and/or small molecules are applied, and then washed with increasing stringency. Since each surface/stringency combination leads to a different adsorption profile, the technique is claimed to provide a means for analysis of a complex mixture. We can extend this approach exponentially, by generating thousands of different surface chemistries on Nanobar Codes. Each chemistry will be identified by its Nanobar Code ID; subsequent analysis by surface-assisted laser desorption

ionization (SALDI) (17) will lead to a vastly improved analysis of complex mixtures. We have already demonstrated the ability to analyze proteins isolated by affinity to derivitized Nanobar Code tags using both MALDI and ESI mass spectrometry (see Figure 5).



(a)



$C_6H_{13}NO_5$

Mol. Wt.: 179.17

m/z

(b)

Figure 5. Two mass spectra demonstrating the capabilities of Nanobar Code Identification Tags™ to fish out proteins in human plasma. (a) MALDI mass spectrogram of 1pg/ml of insulin bound to Nanobar Codes coated with antibody to insulin. (b) ESI mass spectrogram of ionically-affinity captured glucosamine by mercaptoundecanoic acid, a self-assembled monolayer (SAM), coated Nanobar Codes.

D. Research Design and Methods

The underlying goal of this work is to create a scalable bioinformatics infrastructure that can be used to efficiently collect, process, store and mine mass spectrometer data, assay conditions as well as all sample separation procedures. The software tools that are developed should be simple to use, integrate well with our present infrastructure (see Figure 6), and be designed to support SurroMed's mission of biomarker discovery as well as public use of some of the tools. The tasks for such an implementation include:

- Integration of mass spectrometry data from multiple commercial platforms into a unified format using tools that require minimal user intervention;
- Investigation and derivation of algorithms to reduce noise and extract relevant peaks irrespective of which mass spectrometer generated the data;
- Generation of an Oracle database schema to support information on Nanobar Code-Identification Tag based protein extraction, serum separations, experimental designs, and mass spectrometry data;
- Linking of current Nanobar Code imaging software to Oracle database;
- Investigation and implementation of either neural networks or clustering algorithms to distinguish patient mass spectra for differential phenotyping;
- Incorporation of either vendor or custom visualization tools to assist chemists in mass spectrometry-based assay development.

Below research designs for each of these proposed tasks is presented.

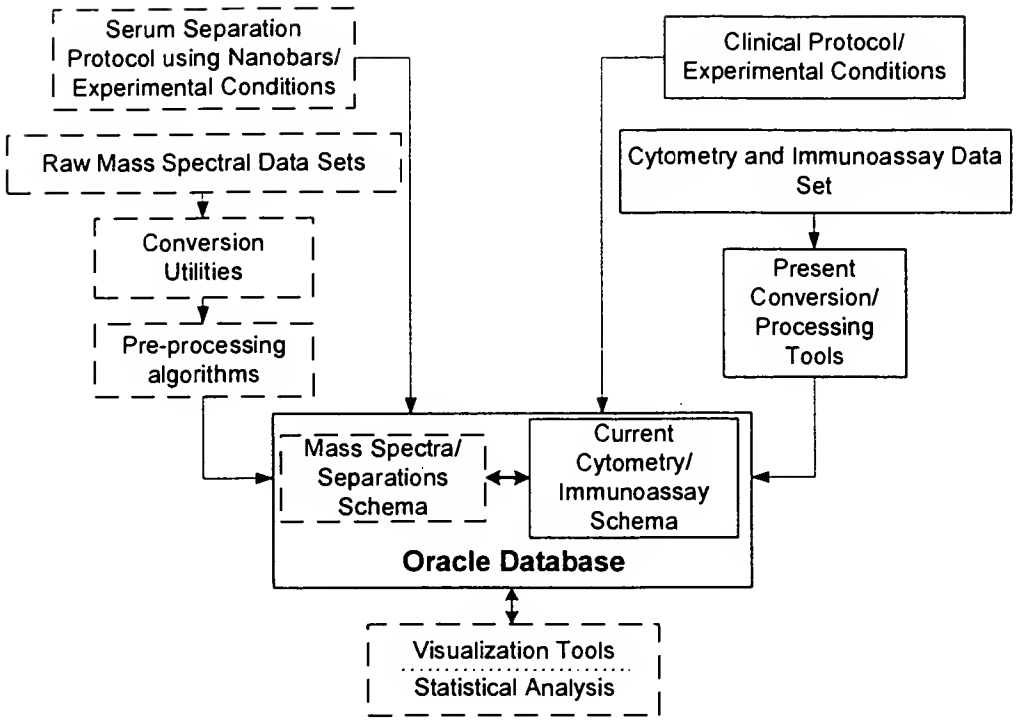


Figure 6. SurroMed's present and proposed informatics infrastructure. The dashed-boxes represent proposed informatic-based projects. The solid-boxes on the right are presently in place.

Unifying mass spectrometry data formats

Although we have already begun software development to unify data formats, this area remains a challenge since most mass spectrometry vendors do not provide open access to their data formats. Instead, these vendors try to package conversion programs into their own software packages. This unfortunately leads to a plethora of independent software tools that must be used sequentially to move data from one form to another before processing. As indicated in the preliminary studies section, we have collaboration with Bear Instruments Inc. to convert all other mass spectrometer data formats to their own scalable data format. This is due to the lack of agreement among mass spectrometer vendors on a unified format. One such attempt at unification resulted in the NetCDF (network common data form) (18) file format. However, while some vendors allow for conversion to this format, the inability of this format to support many of the operating modes of different instruments limits its use.

Our plan is to customize conversion using a combination of component object modules (COM) that are usually supplied by a vendor, along with our own custom software programming

expertise. We will integrate conversion, processing, and upload to Oracle into a single software package. This software will be written in Microsoft Visual C++ in order to maximize conversion and processing speed. In addition, we propose to make the data formats and conversion tools we create publicly available by making them freely downloadable from our web site.

Mass spectrometry algorithms to reduce noise and extract peaks

Our initial purpose for investigating mass spectrometric data processing algorithms is to reduce data size. A typical half hour run on an LC/MS system produces approximately 15MB of data. Since it is anticipated that less than 10 protein or organic molecule components will be found per fraction (see *Separations* in background and significance), and since each protein will be represented by fewer than 10 peaks, the whole 15MB of data could be reduced to about 100 mass-intensity pairs. The challenge is to determine which 100 peaks are significant.

We have begun testing of some common peak extraction algorithms, namely CODA, BB, and Backfolding, to determine their reduction potential and accuracy. We have noticed many weaknesses in these algorithms, especially for differential phenotyping applications. We have recognized the non-normal distribution of noise in the mass spectrograms and have begun testing of nonlinear filters including a median filter that initial testing shows promise.

Many of these algorithms and the mining algorithms we describe later are computationally intensive, taking between several minutes and several hours per data set on our current system. These would perform faster on a dedicated compute server. We propose to purchase a high performance UNIX 4-processor workstation to speed performance.

Our plan for testing these algorithms and future custom algorithms requires the generation of multiple data sets. It is important, from an algorithmic standpoint, to understand the contributions of noise from multiple sources in our separation procedure and analysis, which is shown in Figure 7. To identify and quantify the various noise contributions within the final mass spectrogram, we have begun variability testing. This testing will include: multiple measurements of the same biological sample; multiple separations on the same sample; and multiple runs on the mass spectrogram with different users. These tests will include serum samples with and without spiked proteins. By titrating the spiked proteins, we will better understand both the sensitivity limits of our instrument and the capability of our pre-processing algorithms to distinguish noise from signal.

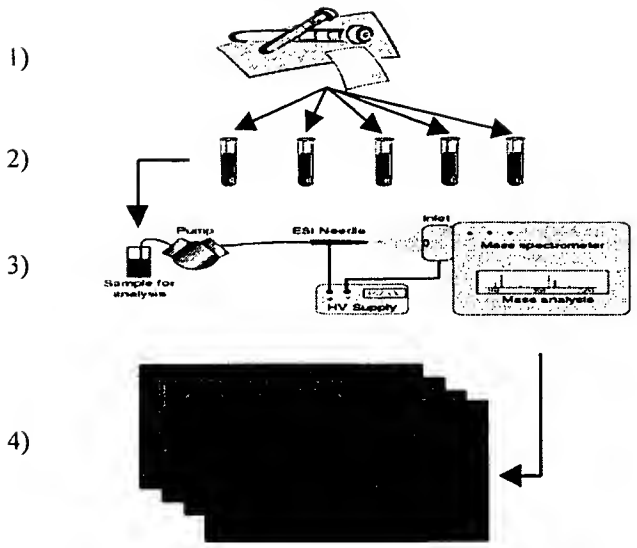


Figure 7: From the analysis standpoint, variation can occur at every step of the measurement process. 1) Biological variation, the variation of interest to SurroMed, (i.e. biomarkers); 2) variations in the separations procedure; 3) mass spectrometric instrument noise; and 4) variability in operation leading to mass spectra.

Development of an Oracle database schema

The development of an Oracle database schema will progress in multiple stages. The database schema will be similar to that of our existing schemas for cytometry and immunoassay experiments described in the preliminary studies section. Since we are in an assay development stage, whereby our intent is to refine the procedures used for separation and sample preparation (as well as defining optimal mass spectrometric instrument settings), our initial database tables pertaining to sample preparation will be quite flexible. This will allow for changes and refinements of our assay procedures.

The format of mass spectral data offers a unique challenge in terms of database storage. Since the number of mass-intensity pairs can be quite variable between assays, we will investigate the optimal method to store variable length lists in the Oracle relational database.

The key will be finding a method that is not just efficient in terms of storage, but also rapid in terms of access for searching/mining the lists.

Since Oracle is a relational database, the key to effective schema design is optimizing the relationships between the tables. Performance was increased more than ten fold by optimization of our cytometry schema. Time-consuming queries can severely limit our storage and mining capabilities. Migration from our current NT Oracle server to a high speed UNIX Oracle server will not only improve database performance, but also provide us with a more stable platform.

Integration of Nanobar Code imaging software

Second generation Nanobar Code tag readers will automate the detection and readout with custom instrument designs and sophisticated image analysis software that will be capable of detecting and reading each Nanobar Code, and quantifying the fluorescence from molecules bound to the Nanobar. Additionally, this detection system will be designed to allow for highly focused laser excitation of the appropriate wavelength to enable laser-induced desorption of non-covalently bound molecules from the surface of each individual Nanobar for MALDI-TOF or, in the case of LC and/or CE based separation, trigger the mass spectrometer to operate under pre-specified conditions for optimal sensitivity. Custom image analysis software will be developed that is capable, in real time, of locating, decoding the stripe pattern, and quantifying fluorescence from each Nanobar in the field of view. In addition, it will be utilized to direct a highly focused laser spot, with the spot size matched to the largest dimension of the Nanobar, at each individual Nanobar. Beam steering optics can be used to sequentially illuminate and desorb proteins and/or molecules from each Nanobar, which will then be accelerated into a time-of-flight mass spectrometer. Individual Nanobars can be chosen for mass spectral analysis based on the fluorescence signal detected, thus minimizing measurement time by analyzing only those rods with significant binding.

A second scenario to combine Nanobar Codes and mass spectrometry involves the use of microfluidics linked to the ESI-based mass spectrometer. We have been actively investigating the microfluidics market for potential technologies that will be compatible with our Nanobar Code Identification Tag technology. SurroMed has already undertaken a partnership with an expert in the field of fringe flow cytometry (19) who has investigated large striated particles such as chromosomes in flow. We have begun assembling a flow cytometer compatible with Nanobar

Codes. The recognition of Nanobar Codes will involve flowing the tags through an interference pattern of light. The variability in metallic reflectance of the identification tag flowing through a spatially varying intensity light field will produce a characteristic light signal. We will need to process this signal in order to determine the Nanobar Code flavor. We intend to write recognition software such that the recognition will trigger decisions in terms of the Nanobar Code tag's flow path (using microfluidic valves). The recognition would trigger a flow path into an ESI mass spectrometer. The mass spectral data would then be merged with the recognition data and both would be sent to the Oracle database. New algorithms, software, and database submodels will need to be developed to accommodate for these new experiments.

Multivariate statistical analysis of mass spectrometer data

Multivariate statistical approaches encompass a suite of tools that are used to discriminate and classify data sets. We have experience in these approaches from our work in analyzing cytometry and immunoassay data. Fisher Discriminant Analysis (1), Support Vector Machines (SVM) (20) and Principal component analysis (PCA) (12) are a few of the methodologies we have either implemented and/or considered. Mass spectral data poses new and interesting challenges with respect to applying these methods. Because of the non-linearities built into the mass spectral system, we intend to add neural networks to our repertoire of tests. Neural networks are a standard approach to modeling nonlinear systems.

In terms of challenges to these analyses, we foresee that our results will be highly dependent on our choice of mass spectral preprocessing algorithms. We will need to strike a delicate balance in terms of sensitivity. As an example, consider principal component analysis. PCA attempts to find the fewest components (variables) that can describe the inherent complexity of an original data set. When one maps a complete data set as a function of some minimal number of components, clusters of data points can be found that represent "similar" data sets. Unfortunately, while this analysis might be excellent for differentiating mass spectra representing a number of completely disparate compounds, it would probably suffer when comparing spectra that differ by say, one or two mass-intensity peaks. This is an area that we need to investigate and develop new algorithms.

We expect that, even after reduction of the MS data, the amount of measurement data to be mined will remain very large when compared to the number of blood samples available to the

study. In order to find accurate biomarkers, a potentially large number of measurement combinations must be considered and analyzed. The challenge is to develop data mining algorithms that can model clinical endpoints with acceptable predictive accuracy without suffering the high combinatorics inherent to the problem.

Visualization tools in assay development and mining

Due to the large data content involved in mass spectrometry, it is important to generate visual tools to supplement and hasten assay development. The type of feedback useful to a chemist will be an image that eliminates unnecessary noise, and that provides only those spectral peaks of interest. Since mass spectral data run on an LC/MS instrument consists of essentially 3-D data sets (the variables being retention time, m/z, and intensity), the data load can get quite large (30Mb/hour). With multiple assays, viewing many 30Mb data files can surpass the capabilities of the computer system. Thus, a reduction of this 3-D data set is essential before viewing.

Most commercial software tools offer many display options for mass spectrometric data. However, they were not developed around the task of assay development. We will generate a custom visualization interface that combines pre-processing steps (to reduce data size) with the ability to view data integrated among either m/z or retention time (in ESI/MS data sets). More importantly, we will enhance the comparison tools between mass spectrograms. These tools include the ability to subtract (intensities), correlate (intensities, time and m/z), translate (time), and scale (time) mass spectrograms. For instance, the translation value required to overlap two TICs along the retention time axis from two different assay conditions, could be useful for adjusting chromatographic conditions.

Visualisation tools for the MALDI data will involve comparing different flavors of Nanobar Codes to select/screen them for resolving overlapping/common components extracted from the sample. Later this will be extended to differential analysis of normal versus disease samples with the common flavors of Nanobar Codes. An envisioned interface would be one that would plot flavors of Nanobar Code tag chemistries (~100 - 1000) on one axis and m/z on another with the intensity shown on a gray scale. Similar interfaces for multidimensional LC, rather than Nanobar Codes, have been developed by other groups and have been found to be very beneficial to assay development (21).

Summary

Given our current ability to use derivatized Nanobar Code Identification Tags™ as a separation method that we have coupled to mass spectrometry for identification and quantification of both proteins and small molecules in complex samples, SurroMed is on the verge of being able to use this unique combination of technologies as another component in our differential phenotyping platform. The development of an informatics platform to integrate all of these tools has begun, but must be extended to include tools to analyze, store and mine the data we are now able to collect. When combined with our cytometry and immunoassay platforms, a very powerful suite of tools for biomarker discovery will be generated.

E. Human Subjects

Not Applicable.

F. Vertebrate Animals

Not Applicable.

G. Literature Cited

1. Afifi, A.A; Clark, V., *Computer-Aided Multivariate Analysis*, Chapman & Hall: London, **1996.**

2. Hochberg Y., "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika*, **1988**; 75, 800-803.

3. Biller, J.E.; Biemann, K. "Reconstructed Mass Spectra, A Novel Approach for the Utilization of Gas Chromatograph-Mass Spectrometer Data," *Anal. Lett.*, **1974**, 7, 515-528.

4. Colby, B.N. "Spectral Deconvolution for Overlapping GC/MS Components," *J. Am. Soc. Mass. Spectrom.*, **1992**, 3, 558-562.

5. Fleming, C.M.; *et. al.* "Windowed Mass Selection Method: A New Data Processing Algorithm for Liquid Chromatography-Mass Spectrometry Data," *J. Chromatogr. A*, **1999**, 849, 71-85.

6. (a) *Applied Numerical Linear Algebra*, James W. Demmel, Society for Industrial and Applied Mathematics, **1997.** (b) Iwata, T.; Koshoubu, J. "Minimization of Noise in Spectral Data," *Appl. Spectrosc.* **1996**, 50, 747-752. (c) Iwata, T.; Koshoubu, J. "New Method to Eliminate the Background Noise from a Line Spectrum," *Appl. Spectrosc.*, **1994,** 48, 1453-1456. (d) Younan, N.H; Zhang, H. "Adaptive Signal Enhancement of Laser-Induced Breakdown Spectroscopy" *Appl. Spectrosc.*, **1999**, 53, 612-617.

7. Windig, W.; Phalp, J.M; Payne, A.W. "A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry," *Anal. Chem.*, **1996,** 68, 3602-3606.

8. Muddiman, D.C. *et. al* "Application of Sequential Paired Covariance to Capillary Electrophoresis Electrospray-Ionization Time-of-Flight Mass Spectrometry: Unraveling the Signal from the Noise in the Electropherogram," *Anal. Chem.*, **1995**, 67, 4371-4375.

9 Muddiman, D.C. *et. al.* "Application of Sequential Paired Covariance to Liquid Chromatography - Mass Spectrometry Data: Enhancements in both the Signal-to-Noise Ratio and the Resolution of Analyte Peaks in the Chromatogram," *J. Chromatogr. A*, **1997**, 771, 1-7.

10. (a) Pool, W.G.; de Leeuw, J.W.; van de Graaf, B. "Backfolding Applied to Differential Gass Chromatography/Mass Spectrometry as a Mathematical Enhancement of Chromatographic Resolution," *J. Mass Spectrom.*, **1996**, 31, 509-516. (b) Pool, W.G.; de Leeuw, J.W.;van de Graaf, B. "Automated Extraction of Pure Mass Spectra from Gas Chromatographic/Mass Chromatographic Data," *J. Mass Spectrom.*, **1997**, 32, 438-443.

11. Lee, T.A.; Headley, L.M.; Hardy, J.K. "Noise Reduction of Gas Chromatography/Mass Spectrometry Using Principal Component Analysis," *Anal. Chem.*, **1991**, 63, 357-360.

12. Mavrovouniotis, M.L; Harper, A.M.; Ifarraguerri, A.I. "Classification of Pyrolysis Mass Spectra of Biological Materials Using Convex Cones," *J. Chemom.*, **1994**, 8, 305-331.

13. (a) Lohninger, H.; Stanel, F. "Comparing the Performance of Neural Networks to well-established Methods of Multivariate Data Analysis: the Classification of Mass Spectral Data," *Fresenius' J. Anal. Chem.*, **1992**, 344, 186-189. (b) Wan, C.; Harrington, P.B. "Self-Configuring Radial Basis Function Neural Networks for Chemical Pattern Recognition," *J. Chem. Inf. Comput. Sci.*, **1999**, 39, 1049-1056. (c) Goodacre, R. "Use of Pyrolysis Mass Spectrometry with Supervised Learning for the Assessment of the Adulteration of Milk of Different Species," *Appl. Spectrosc.*, **1997**, 8, 1144-1153.

14. Martin, B. *et. al.*"Orthogonal Self-Assembly of Colloidal Multimetal Nanorods," *Adv. Mater.*, **1999**, 11, 1021-1025.

15. *Applications of Solid Phase Microextraction*, Pawliszyn J., ed., The Royal society of Chemistry, Cambridge, UK, **1999**.

16. Hutchens, T.W. *et al.*, "New Desorption Strategies for the Mass Spectrometric Analysis of Macromolecules," *Rapid Commun. Mass Spectrom.* **1993**, 7, 576.

17. Sunner, Jan; Dratz, Edward; Chen, Yu-Chie. "Graphite surface-assisted laser desorption/ionization time-of-flight mass spectrometry of peptides and proteins from liquid solutions," *Anal. Chem.*, **1995**, 67, 4335-4342.

18. http://www.unidata.ucar.edu/packages/netcdf/

19. Mullikin, J., *et al.*," Fringe-scan flow cytometry," **1988**. *Cytometry*, 9, 111-120.

20. Cherkassky, V; Mulier, F., *Learning From Data*; John Wiley & Sons, Inc.: New York, **1998**.

21. Wall, D.B. *et al.*, "Isolectric Focusing Nonporous RP HPLC: A Two-Dimensional Liquid-Phase Separation Method for Mapping of Cellular Proteins with Identification Using MALDI-TOF Mass Spectrometry," *Anal. Chem.*, **2000**, 72, 1099-1111.

S:\ClientFolders\Surromed\34\provisional.doc

NONLINEAR FILTER FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA

FIELD OF THE INVENTION

[0001] This invention relates generally to analysis of data collected by analytical techniques such as chromatography and spectrometry. More particularly, it relates to a nonlinear filter for noise reduction in mass chromatograms acquired by liquid chromatography-mass spectrometry.

BACKGROUND OF THE INVENTION

[0002] Liquid chromatography-mass spectrometry (LC-MS) is a well-known combined analytical technique for separation and identification of chemical mixtures. Chromatography separates the mixture into its constituent components, and mass spectrometry further analyzes the separated components for identification purposes.

[0003] In its basic form, chromatography involves passing a mixture dissolved in a mobile phase over a stationary phase that interacts differently with different mixture constituents. Components that interact more strongly with the stationary phase move more slowly and therefore exit the stationary phase at a later time than components that interact more strongly with the mobile phase, providing for component separation. A detector records a property of the exiting species to yield a time-dependent plot of the property, e.g., mass or concentration, allowing for quantification and, in some cases, identification of the species. For example, an ultraviolet (UV) detector measures the UV absorbance of the exiting analytes over time. When liquid chromatography is coupled to mass spectrometry, mass spectra of the eluting components are obtained at regular time intervals for use in identifying the mixture components. Mass spectra plot the abundance of ions of varying mass-to-charge ratio produced by ionizing and/or fragmenting the eluted components. The spectra can be compared with existing spectral libraries or otherwise analyzed to determine the chemical structure of the component or components. Note that LC-MS data are two-dimensional; that is, a discrete data point (intensity) is

obtained for varying values of two independent variables, retention time and mass-to-charge ratio (m/z).

[0004] LC-MS data are typically reported by the instrument as a total ion current (TIC) chromatogram, the sum of all detected ions at each scan time. Peaks in the chromatogram represent separated components of the mixture eluting at different retention times. A noise-free chromatogram **10**, shown in FIG. 1A, appears as a series of smooth peaks **12a-12c**, each extending over multiple scan times. As shown in the TIC chromatogram of FIG. 1B, however, LC-MS data often have high-intensity noise spikes **14a-14d** superimposed on the peaks. Noise spikes typically do not extend beyond one scan time. If the TIC chromatogram has little noise, an operator can determine the total number of peaks and then examine each peak's corresponding mass spectrum to identify the eluted species. However, as the amount of noise present increases, it becomes more difficult for the operator to distinguish the chromatographic peaks, particularly if the noise level is higher than the signal level. In such cases, the operator is left to manually examine each individual mass spectrum, select the mass-to-charge ratios corresponding to known or likely mixture components, and then assemble a reduced total ion current chromatogram from the selected masses only. Such a procedure is clearly very time consuming. Furthermore, when the mixture contains unknown analytes, the operator cannot confidently determine which mass spectral peaks are noise and which are actual peaks. Thus the only recourse the operator has is to adjust various instrument parameters and repeat the experiment with a different sample, hoping for less noise in the resulting chromatogram.

[0005] Recently, LC-MS has been used to analyze complex biological mixtures. Proteomics is a relatively new field that aims to detect, identify, and quantify proteins to obtain biologically relevant information. Both proteomics and metabolomics (the detection, identification, and quantitation of metabolites and other small molecules such as lipids and carbohydrates) may facilitate disease mechanism elucidation, early detection of disease, and evaluation of treatment. Recent advances in mass spectrometry have made it an excellent tool for structural determination of proteins, peptides, and other biological

molecules. However, proteomics and small molecule studies typically have a set of requirements that cannot be met by manual interpretation of the LC-MS data.

[0006] First, these studies require high-throughput analysis of small volumes of biological fluid. Manual data interpretation creates a bottleneck in sample processing that severely limits the number of samples that can be analyzed. While large available sample volumes allow an operator to adjust parameters by trial and error to obtain adequate chromatograms and spectra, biological samples are available in such small volumes that it is imperative to extract useful information from any available sample. Second, unlike traditional research applications, in which a relatively small amount of data is required, the paradigm of these studies is to acquire enormous amounts of data and then mine the data for new correlations and patterns. Manual data analysis is therefore unfeasible. In addition, biological samples are generally mixtures of unknown compounds, often at very low concentrations, and so it is not desirable to extract only known spectra and discard the remaining data, as it would be for studies involving quantification of known compounds in a mixture. There are also generally substantially more peaks in biological samples than in, for example, samples of environmental pollutants. Finally, LC-MS instruments produce an enormous amount of data: a single one-hour chromatographic run can produce up to 80 MB of binary data. For storage and subsequent data mining purposes, it is highly desirable to reduce the amount of data to retain information while discarding noise. To satisfy these requirements, a data analysis method is needed that can acquire a large amount of data from low-volume biological mixtures, extract useful information from the resulting noisy data set, and identify unknown compounds from the extracted information. An essential component of such a method is the ability to distinguish peaks from noise automatically.

[0007] The component detection algorithm (CODA) is an automated method for selecting mass chromatograms with low noise and low background. CODA is described in W. Windig et al., "A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry," *Anal. Chem.*, 68 (1996): 3602-3606. The method computes a smoothed and mean-subtracted version of each mass

chromatogram, compares it with the original chromatogram, and calculates a similarity index between the two. Chromatograms whose similarity index exceeds a threshold value are retained and combined to form a reduced total ion chromatogram, while other chromatograms are rejected. CODA has proven very effective at selecting high-quality mass chromatograms. However, it can only accept or reject entire chromatograms based on their noise level, but cannot filter noise from an individual chromatogram. As a result, noisy chromatograms that contain useful information are eliminated, and important peaks may not be detected.

[0008] Techniques exist for filtering noise and background from spectrometric data. For example, U.S. Patent No. 5,995,989, issued to Gedcke et al., describes a filtering method in which an average background level and an average deviation from the background are computed and used to define a local threshold for each data point. Points exceeding the threshold are retained, while points below the threshold are considered to be noise and discarded. The technique described in Gedcke et al. is only effective for noise levels that are substantially below the level of the peaks. For data such as that illustrated in FIG. 1B, high-intensity noise spikes cannot be removed using the disclosed method.

[0009] In U.S. Patent No. 6,112,161, issued to Dryden et al., a method for enhanced integration of chromatography or spectrometry signals is described. A baseline signal is computed from a moving average of the actual signal. The difference between the baseline and actual signal is a baseline-adjusted signal containing peaks and high-frequency noise. An intensity range of the noise is determined, and all signal outside of this range is considered to be peaks, while signal inside this range is considered to be noise. As with the method of Gedcke et al., the method of Dryden et al. can only be used when the noise intensity is substantially lower than the signal intensity. Because LC-MS data often has noise values exceeding the signal values, the method of Dryden et al. is not effective at removing noise from LC-MS data.

[0010] A moving median digital filter has been used to remove noise from mass spectrometry and potentiometric titration data, as described in C.L. do Lago et al.,

"Applying moving median digital filter to mass spectrometry and potentiometric titration," *Anal. Chim. Acta*, 310 (1995): 281-288. Each data point is replaced by the median of the values in a window surrounding the point. With respect to the mass spectrometry data, the filter is applied both to the electron multiplier output, i.e., the ion abundance values, and to the magnetic field sensor, i.e., the mass-to-charge ratio. The method is not, however, applied to two-dimensional data such as LC-MS data. In most cases, state-of-the-art LC-MS instruments do not report the mass spectra as continuous smooth peaks, but rather as centroided data, i.e., single-mass peaks at the average mass value of the true peak. Without centroiding, an unmanageable amount of data would be generated for each spectrum. A moving median filter applied to centroided mass spectral data would remove peaks and noise equally. Because the peak shape is removed in the reported data, filtering or analytical methods cannot be applied to the mass spectra. Moreover, in some cases, one major source of noise, detector noise, can corrupt an entire mass spectrum. If a high fraction of the points in the filter window are corrupted, then a median filter applied to the spectrum cannot remove this noise.

[0011] There is still a need, therefore, for a method for removing noise, particularly high-intensity spikes, from chromatographic and spectrometric data such as LC-MS data.

SUMMARY OF THE INVENTION

[0012] The present invention provides a method for filtering noisy mass chromatograms in two-dimensional liquid chromatography-mass spectrometry (LC-MS) data. The method can remove noise spikes that have a higher intensity than peaks corresponding to eluted analytes, while essentially retaining the peak intensity and shape. It therefore performs substantially better than conventional linear filters that assume a normal distribution of the noise.

[0013] In a method of the invention for characterizing a chemical or biological sample, a series of mass spectra are generated by chromatography (e.g., liquid chromatography) and mass spectrometry. A total ion current (TIC) chromatogram is

obtained from the mass spectra, and a nonlinear filter is applied to the chromatogram, resulting in a filtered total ion chromatogram with lower noise than the original chromatogram. Preferably, individual chromatograms are generated from the series of mass spectra, and the filter is applied separately to each individual chromatogram. The total ion current chromatogram is then reconstructed from the individual filtered chromatograms. Alternatively, the raw mass spectral data can be compared with the filtered chromatograms and the raw data replaced with the corresponding filtered data if the filtered data has a lower intensity value. The TIC chromatogram can then be assembled from the thresholded raw data. Subsequent to filtering, additional post-acquisition processing steps can be performed, such as applying a component detection algorithm to the filtered data to select relatively noise-free individual chromatograms.

[0014] The nonlinear filter can be any suitable nonlinear filter such as a moving median filter or a wavelet decomposition filter, and the method preferably also includes selecting and optimizing one or more parameters of the nonlinear filter. For example, the parameter can be selected based on the scan rate of the mass spectrometer or on subsequent data analysis, such as peak selection, of the mass spectra.

[0015] Also provided by the present invention is a program storage device accessible by a data analysis machine and tangibly embodying a program of instructions executable by the machine to perform method steps for the above method.

BRIEF DESCRIPTION OF THE FIGURES

[0016] FIGS. 1A-1B are schematic diagrams of a noise-free total ion current chromatogram and a noisy total ion current chromatogram, respectively, as known in the prior art.

[0017] FIGS. 2A-2B illustrate a moving median filter of the invention applied to a chromatogram peak and to a chromatogram noise spike, respectively.

[0018] FIG. 2C illustrates a moving average filter applied to a chromatogram noise spike.

[0019] FIG. 3 is a flow diagram of a nonlinear filter method of the present invention.

[0020] FIG. 4 illustrates the application of a moving median filter with a poorly chosen window size to a chromatogram peak.

[0021] FIGS. 5A-5B show total ion chromatograms, base peak traces, and two-dimensional LC-MS data obtained from an LC-MS experiment of a proteolytic digest of human serum, before and after application of the moving median filter of the invention, respectively.

[0022] FIG. 5C shows the total ion chromatogram, base peak trace, and two-dimensional LC-MS plot of FIG. 5A after application of a mean filter.

[0023] FIG. 6 is a block diagram of a hardware system for implementing the method of FIG. 3.

DETAILED DESCRIPTION OF THE INVENTION

[0024] The present invention provides a method for filtering chromatographic and spectrometric data to reduce noise in individual mass chromatograms, thereby facilitating subsequent selection of peaks or high quality chromatograms for component detection. In liquid chromatography-mass spectrometry (LC-MS) data, the noise intensity is often larger than the intensity of the peaks corresponding to eluted species, making it very difficult to extract meaningful information from the data. The method of the invention is able to reduce substantially such large magnitude noise spikes.

[0025] LC-MS noise originates from a variety of sources corresponding to different components of the system. For example, chemical noise results from column bleed, i.e., long-time elution of strongly-adsorbed species at particular mass-to-charge ratios, low-concentration sample contaminants, and detection of the chromatographic mobile phase. In the mass spectrometer, the ion generation, selection, and detection processes all generate noise. Electronic signal processing and analog-to-digital conversion add white noise to the acquired data. The noise sources and distributions are not well understood for all components, making it difficult to select an appropriate filter.

[0026] In methods of the invention, a nonlinear digital filter is applied to individual mass chromatograms to reduce the noise level. A mass chromatogram is a plot of intensity versus retention time for a particular range of mass-to-charge ratio of detected ions. A nonlinear filter applies a nonlinear function to the data to be filtered. A nonlinear filter is particularly well-suited to LC-MS data because of the noise distribution characteristics of this type of data. As recognized by the present inventor, noise in LC-MS data is not normally distributed, i.e., does not follow a Gaussian distribution. Additionally, empirical study of LC-MS data has revealed that in individual mass chromatograms, noise spikes typically occur over a single scan time only. Standard filtering techniques for LC-MS data use moving average filters, which are linear filters and therefore only effective at removing normally distributed noise.

[0027] A simple nonlinear filter used in the preferred embodiment of the invention is a moving median filter, illustrated in FIGS. 2A-2B. A moving median filter replaces each point with the median of the points in a window of a given size centered on the selected point. For example, a three-point window examines a selected point and the neighboring point on each side of the selected point. Moving median filters are used for noise suppression in image processing but, to the knowledge of the present inventor, have not previously been applied to chromatographic data. FIG. 2A illustrates the application of a three-point moving median filter to a smooth chromatographic peak extending over multiple MS scan times. The top plot is the raw data, and the bottom plot is the filtered data. Points on the peak side slopes are necessarily the median of the three values in the window, and do not change upon application of the filter. The highest point of the peak is replaced by the larger of the two neighboring values. Thus the moving median filter flattens the peak slightly.

[0028] FIG. 2B illustrates the effect of the same three-point moving median filter on a single-point noise spike. The points surrounding the spike change little, if at all, but the noise spike is replaced by the higher of its two adjacent points, i.e., is completely removed. FIGS. 2A-2B highlight the benefits of a moving median filter: it removes high-intensity noise spikes while retaining the sharpness of peak edges. A three-point moving

average filter applied to the same noise spike is illustrated in FIG. 2C. In this case, each point is replaced by the mean of itself and its two surrounding points. The points at the edge of the noise spike are increased in value, while the spike itself is reduced significantly, but is still present. If the noise is of larger magnitude than the actual peaks, then the filtered noise is comparable to the peaks, and the filter is not effective in reducing noise.

[0029] A flow diagram of a method **20** of the invention for reducing noise in LC-MS data is shown in FIG. 3. First, in step **22,** the time-dependent mass spectra are acquired. Next, in step **24,** a mass chromatogram is generated for each integer mass in the entire set of mass spectra. For example, peaks at masses of 1321.7 and 1322.1 are summed and combined into the mass chromatogram for an integer mass of 1322. Alternatively, data points can be combined into mass ranges that do not necessarily correspond to integer values. In step **26,** the nonlinear filter is applied to each mass chromatogram generated in step **24.** The nonlinear filter can be a moving median filter, a wavelet decomposition filter, or any other suitable nonlinear filter. Next, in an optional step **28,** a component selection algorithm such as CODA is applied to the filtered mass chromatograms.

[0030] Finally, the filtered mass chromatograms are combined into a reduced or filtered total ion current chromatogram in step **30.** One method is simply to sum the intensities at each time point. Alternatively, the raw mass spectral data obtained in step **22** can be thresholded using the filtered chromatograms. To do this, each raw data point is compared with its corresponding point in the filtered chromatograms. Recall that the raw data contain points at non-integer values of mass-to-charge ratio, while the filtered chromatograms contain points corresponding to ranges of mass values. If the intensity value of the raw data exceeds the value of the corresponding filtered point, then the original point is replaced by the filtered value. If not, it is retained.

[0031] The method **20** is typically implemented as part of an automated data analysis method for two-dimensional LC-MS. Data filtered according to the present

invention may be subjected to, for example, peak recognition algorithms and structural identification algorithms. It is anticipated that the filtered data can be much more successfully analyzed by subsequent algorithms than can unfiltered data. In fact, one of the problems with the CODA method is that it removes noisy chromatograms, thereby also removing any information contained within the chromatograms. Filtering noise before applying CODA allows more chromatograms to be retained.

[0032] Although the method **20** is best implemented by applying the nonlinear filter to the individual mass chromatograms and then combining the filtered chromatograms into a reduced total ion current chromatogram, the filter alternatively can be applied directly to the original total ion current chromatogram. In individual mass chromatograms, noise spikes typically occur over a single scan time only and are therefore effectively filtered using a nonlinear filter of the invention. In the total ion current chromatogram, however, spikes that occur at different masses but adjacent retention times can effectively merge to extend over multiple scan times and therefore pass the filter.

[0033] The nonlinear filters used in step **26** have parameters that are adjusted to achieve optimal filtering of the signal. The moving median filter, for example, has one parameter, the window size, the number of points over which the median is computed. The optimal window size is determined by a number of factors including the typical peak width and the scan rate. The peak width of LC-MS data varies with column conditions, flow rate, and mobile and stationary phases, among other factors. The window size should not be wider than the typical peak width, or peaks will be significantly distorted. FIG. 4 illustrates the use of a moving median window that is larger than the peak width. As shown, the peak base is approximately five points wide, while the filter window is nine points wide. The peak is essentially removed, and so this filter width is unacceptable. The filter width must be decreased to three points before the peak can survive the filter substantially unchanged.

[0034] In addition, the scan rate determines the density of points in the chromatogram and therefore also affects the optimal window size. If scans are performed

half as frequently as in the chromatogram of FIG. 2A, all else being equal, the peak contains fewer points and therefore a smaller window is needed to retain the peak while eliminating noise spikes. From the point of view of the present invention, therefore, it is desirable to scan more frequently. In a preferred embodiment, the method derives an expected peak width from the chromatography parameters and resolution and then selects a window size based on the expected peak width.

[0035] Additionally, the optimal parameters are determined by the quality of the resulting reduced total ion chromatogram and the ease and accuracy with which subsequent component detection or automated peak picking can be performed.

[0036] In some embodiments, an adaptive window size is employed. The window size is not the same for all data points, but varies based on a number of factors. The window size can be selected based on characteristics of the data by analyzing subsets of points in each mass chromatogram. Alternatively, the variation in window size can be predetermined based on knowledge of the instrument conditions. If peaks at later retention times are known to be broader than peaks at earlier times, then the window is present to increase with retention time.

[0037] Additional nonlinear filters include modified median filters and wavelet decomposition filters. These filters have more parameters than the moving median filter, but their parameters are optimized based on the same principles.

[0038] Methods of the invention preferably use standard algorithms for implementing the various steps. For example, the moving median filter is applied using existing techniques for obtaining the median of a set of points. In one such algorithm, the median window is applied sequentially to the data beginning at the lowest-time data point. Points within the window are ordered and the central point selected as the median. For subsequent points, the earliest-time point is removed and the new point inserted into the correct position in the ordered set. At the edges of the data set, additional points are

appended so that the window can be centered on the first and last points. Preferably, the additional points have the same values as the edge points.

[0039] An example application of the method is shown in FIGS. 5A and 5B. FIG. 5A shows a total ion current chromatogram, base peak trace, and two-dimensional plot acquired from an LC-MS experiment using a proteolytic digest of human serum. The darkness of each point in the two-dimensional plot corresponds to the detected intensity at that mass-to-charge ratio and retention time. Each point in the TIC is the sum of all points directly below it in the two-dimensional plot, while each point in the base peak trace is the maximum value of all points below it. A moving median filter with a window size of 7 points was applied to the mass chromatograms extracted from the data. FIG. 5B shows the resulting filtered data. FIG. 5C shows the results of applying a 7-point mean filter to the same data. Note that the mean filter changes the data very little, while the median filter clearly brings out six smooth peaks.

[0040] Although not limited to any particular hardware configuration, the present invention is typically implemented in software by a system **40,** shown in FIG. 6, containing a computer **42** in communication with an analytical instrument, in this case a LC-MS instrument **44** that includes a liquid chromatography instrument **46** connected to a mass spectrometer **48** by an interface **50.** The computer **42** acquires raw data directly from the instrument **44** via an analog-to-digital converter. Alternatively, the invention can be implemented by a computer in communication with an instrument computer that obtains the raw data. Of course, specific implementation details depend on the format of data supplied by the instrument computer. Preferably, the entire process is automated: the user sets the instrument parameters and injects a sample, the two-dimensional data are acquired, and the data are filtered for subsequent processing or transfer to a suitable database.

[0041] The computer **42** implementing the invention typically contains a processor **52,** memory **54,** data storage device **56,** display **58,** and input device **60.** Methods of the invention are executed by the processor **52** under the direction of

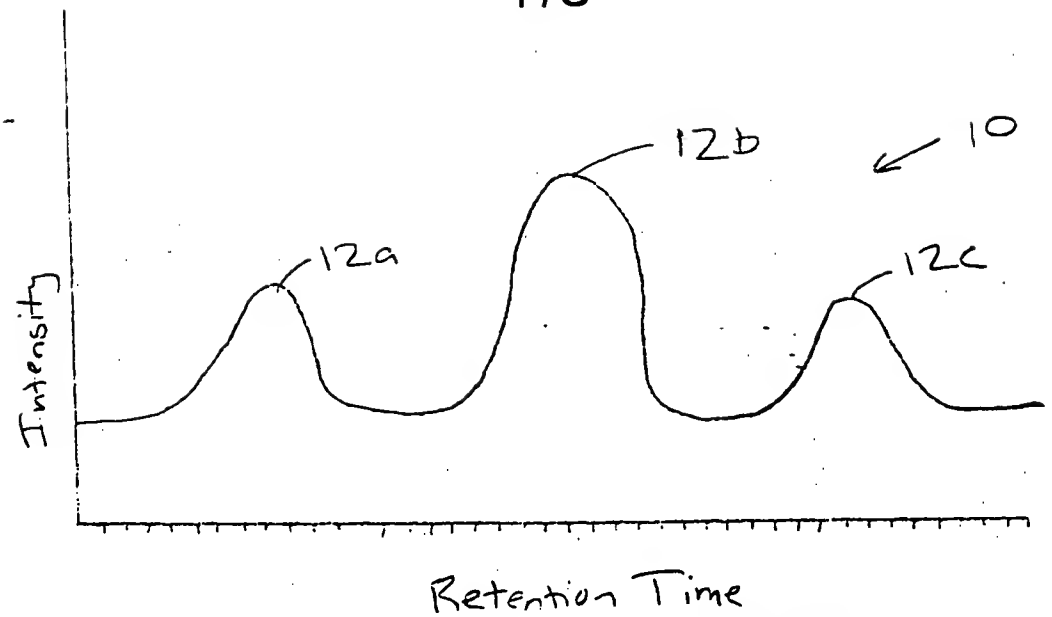
computer program code stored in the computer **42.** Using techniques well known in the computer arts, such code is tangibly embodied within a computer program storage device accessible by the processor **52,** e.g., within system memory **54** or on a computer readable storage medium **56** such as a hard disk or CD-ROM. The methods may be implemented by any means known in the art. For example, any number of computer programming languages, such as Java, C++, or LISP may be used. Furthermore, various programming approaches such as procedural or object oriented may be employed.

[0042] It is to be understood that the steps described above are highly simplified versions of the actual processing performed by the computer **42,** and that methods containing additional steps or rearrangement of the steps described are within the scope of the present invention.

[0043] It should be noted that the foregoing description is only illustrative of the invention. Various alternatives and modifications can be devised by those skilled in the art without departing from the invention. Accordingly, the present invention is intended to embrace all such alternatives, modifications and variances which fall within the scope of the disclosed invention.

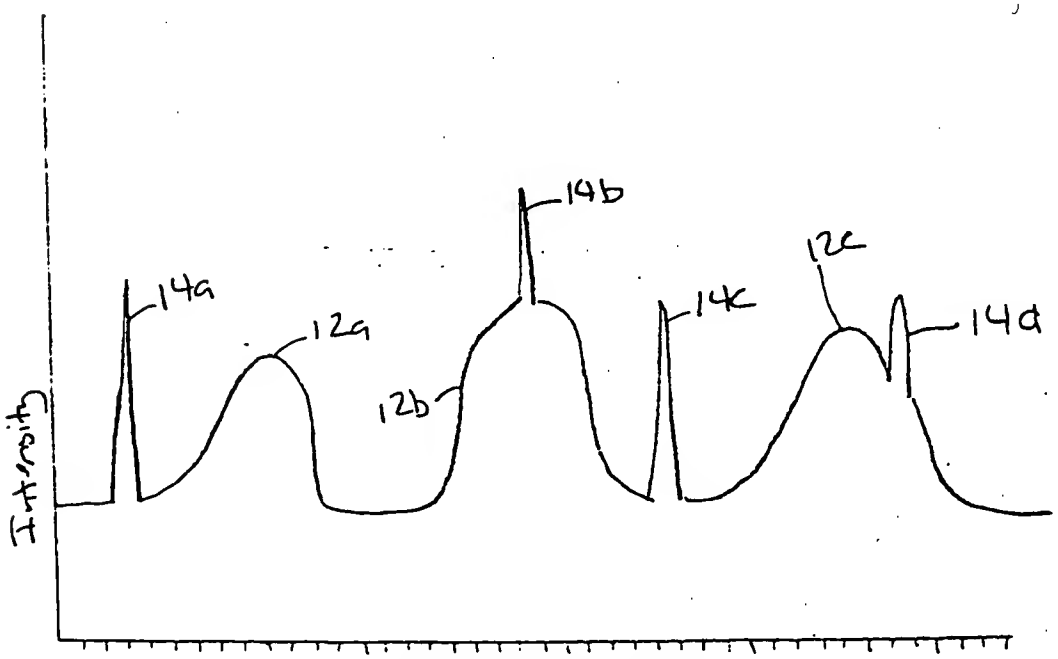
ABSTRACT

High-intensity, spiked noise is reduced in chromatography-mass spectrometry data by applying a nonlinear filter such as a moving median filter to the data. The filter is applied to individual mass chromatograms, plots of ion abundance versus retention time for each detected mass-to-charge ratio, and the filtered chromatograms are combined to form a filtered total ion current chromatogram. Standard linear filters are not effective for reducing noise in liquid chromatography-mass spectrometry (LC-MS) data because they assume a normal distribution of noise. LC-MS noise, however, is not normally distributed.



Prior Art

FIG. 1A



Prior Art

FIG. 1B

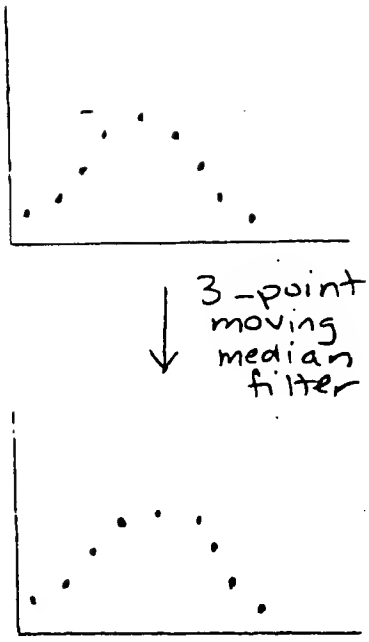


FIG. 2A

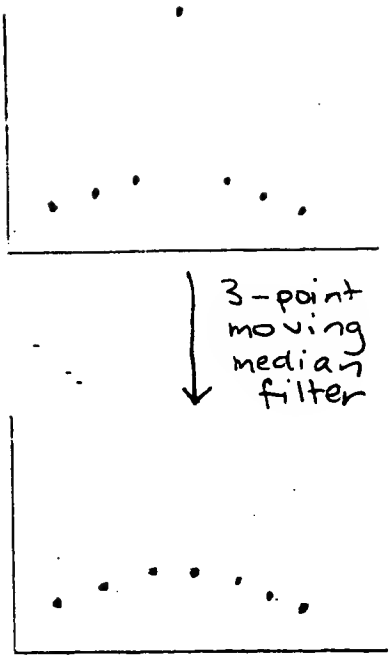


FIG. 2B

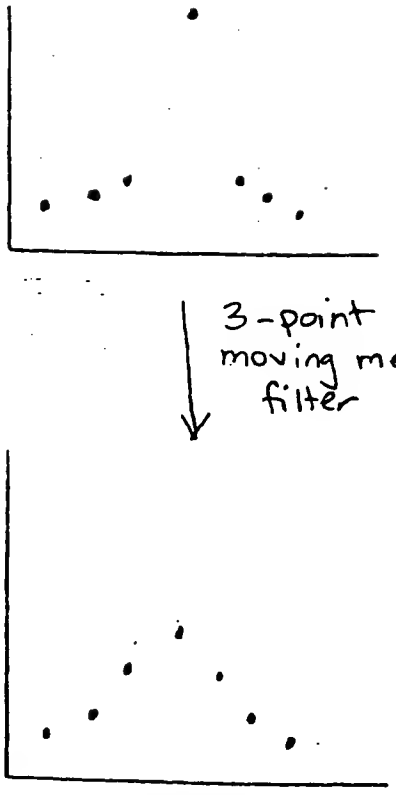


FIG. 2C

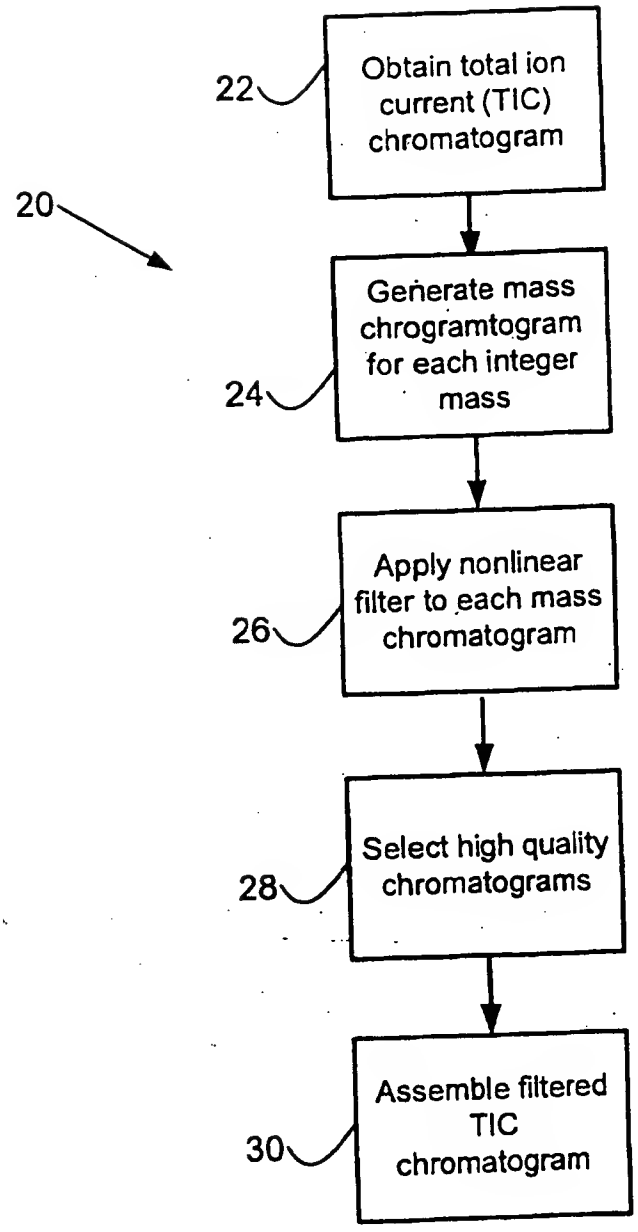
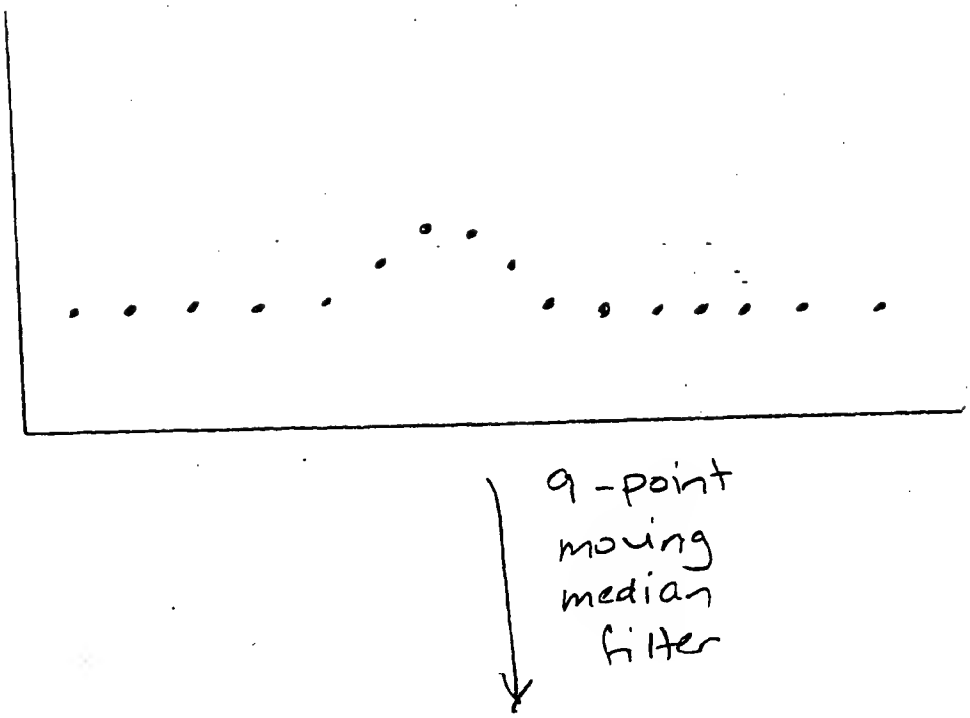

20
22
Obtain total ion current (TIC) chromatogram
24
Generate mass chrogramtogram for each integer mass
26
Apply nonlinear filter to each mass chromatogram
28
Select high quality chromatograms
30
Assemble filtered TIC chromatogram

FIG. 3


9-point
moving
median
filter

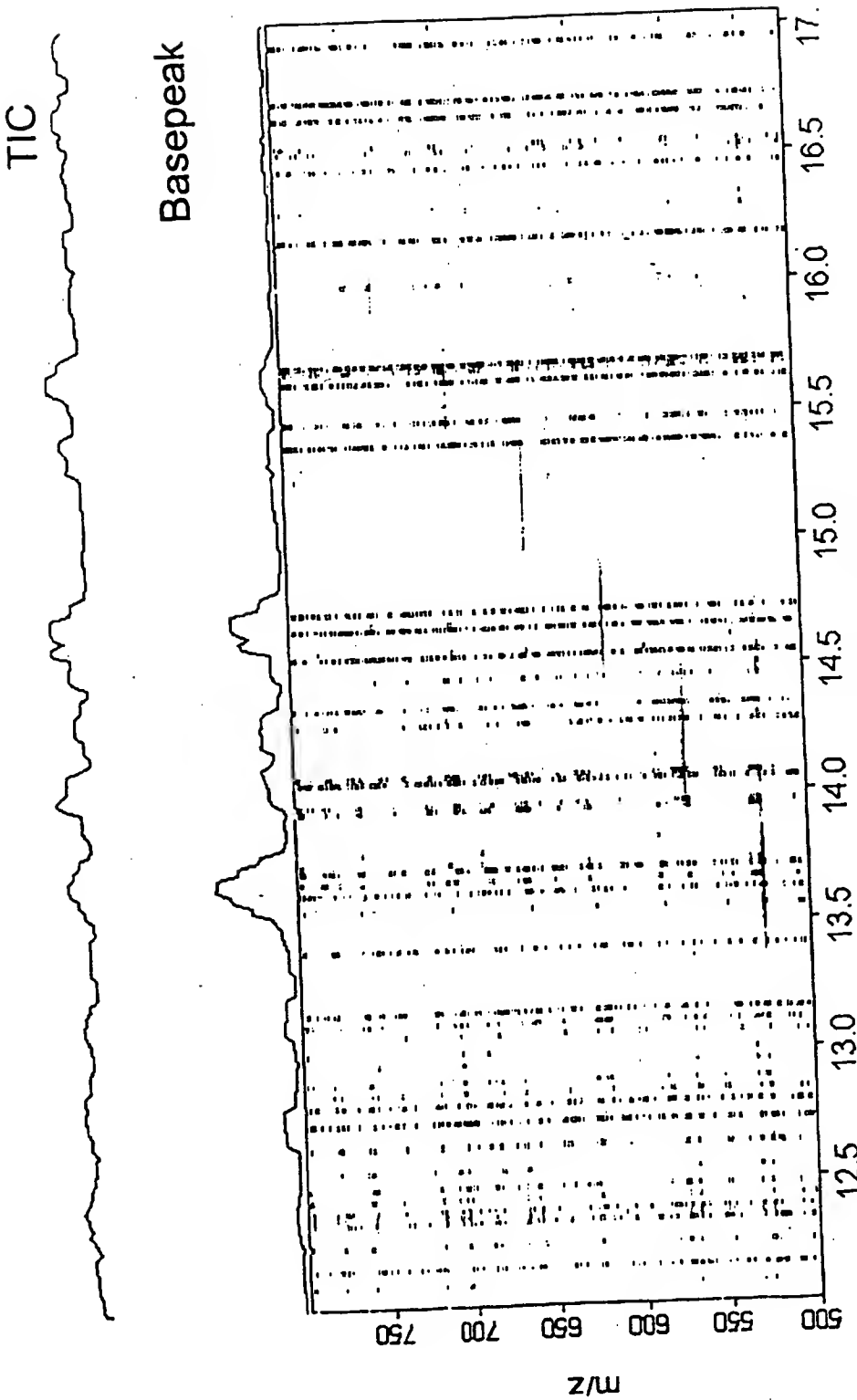

TIC
Basepeak
m/z
500
550
600
650
700
750
Retention Time
12.5
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17.

FIG. 5A

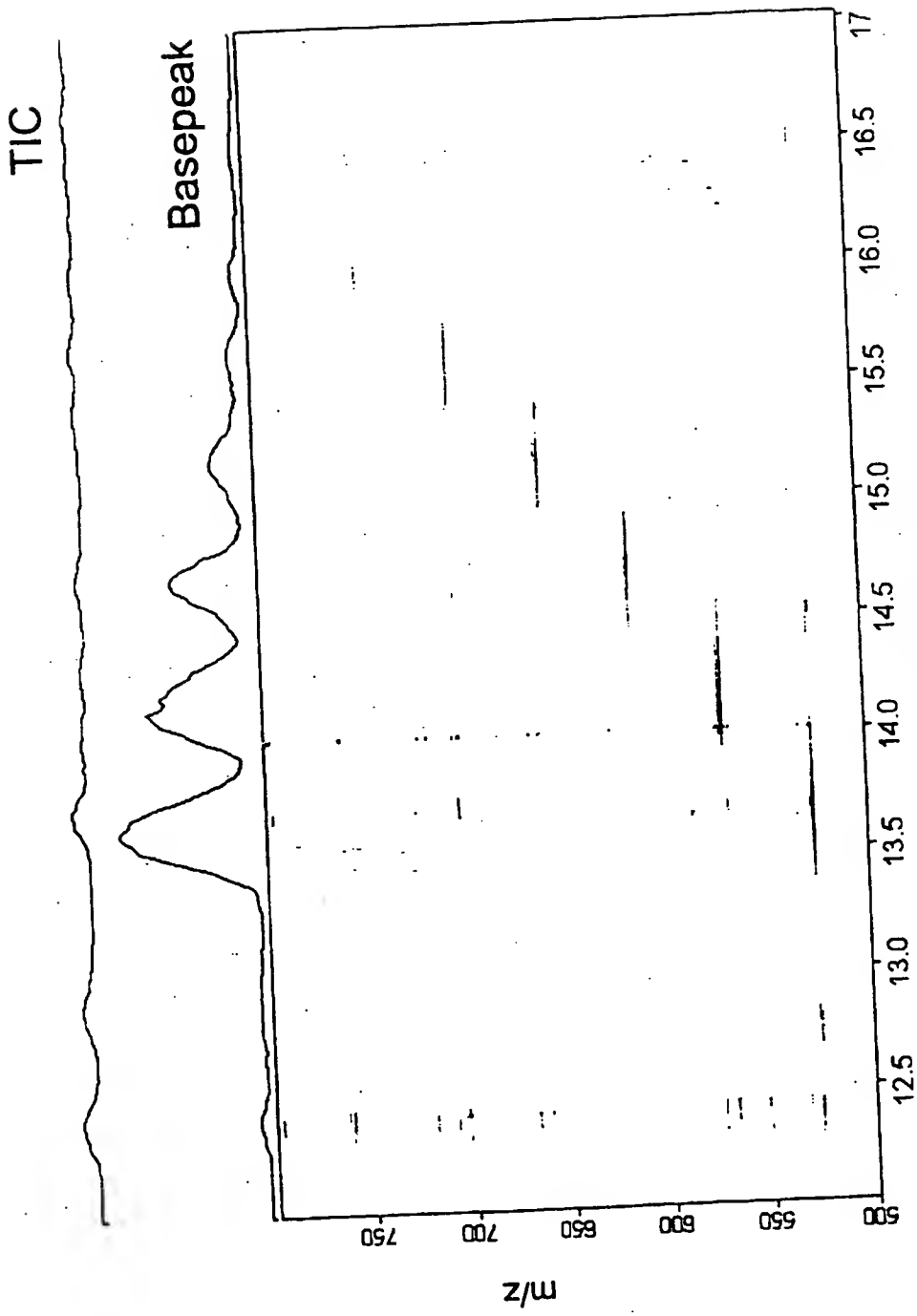

TIC
Basepeak
m/z
750
700
650
600
550
500
Retention Time
12.5
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17

FIG. 5B

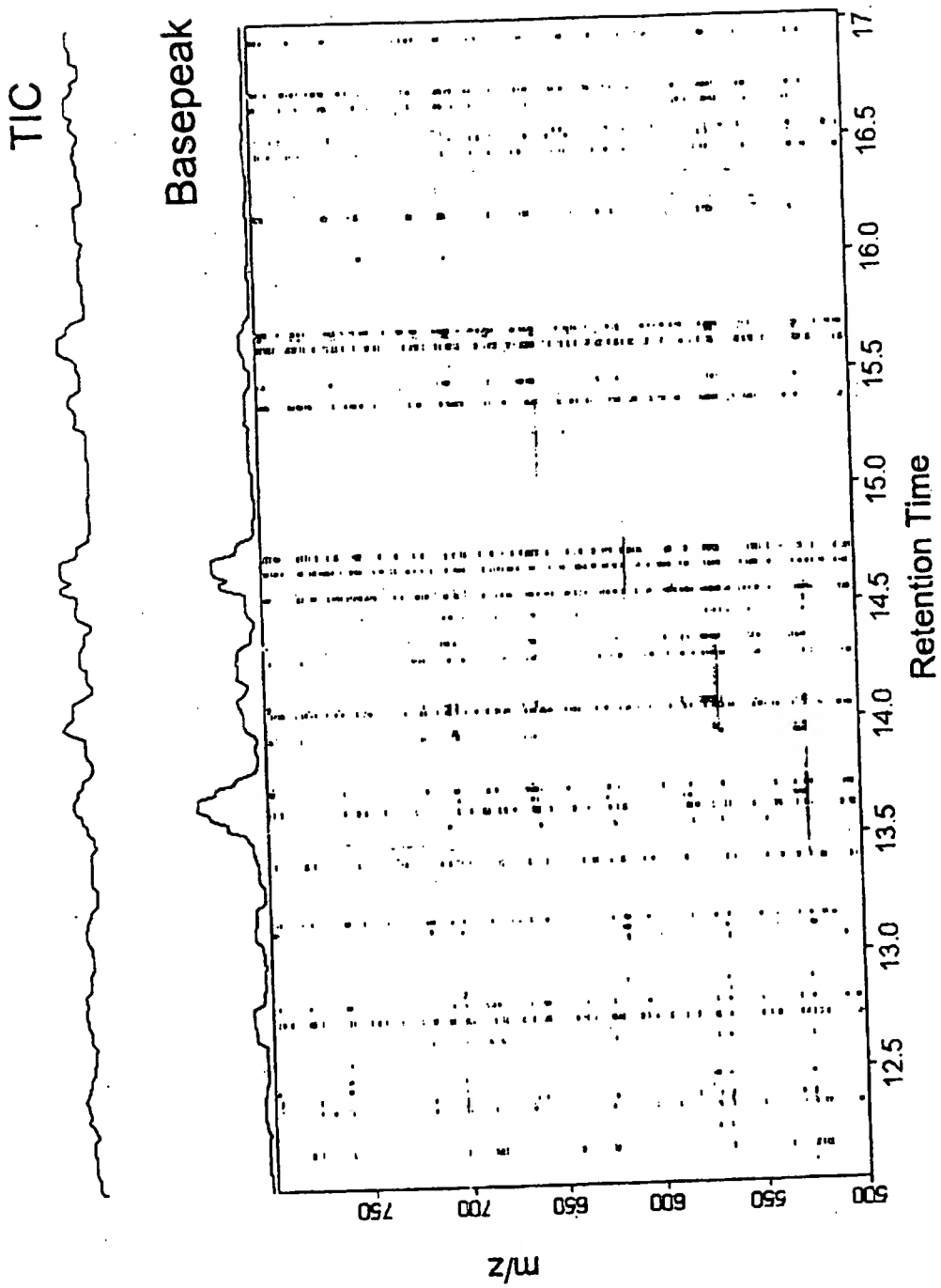

TIC
Basepeak
m/z
500
550
600
650
700
750
Retention Time
12.5
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17

FIG. 5C

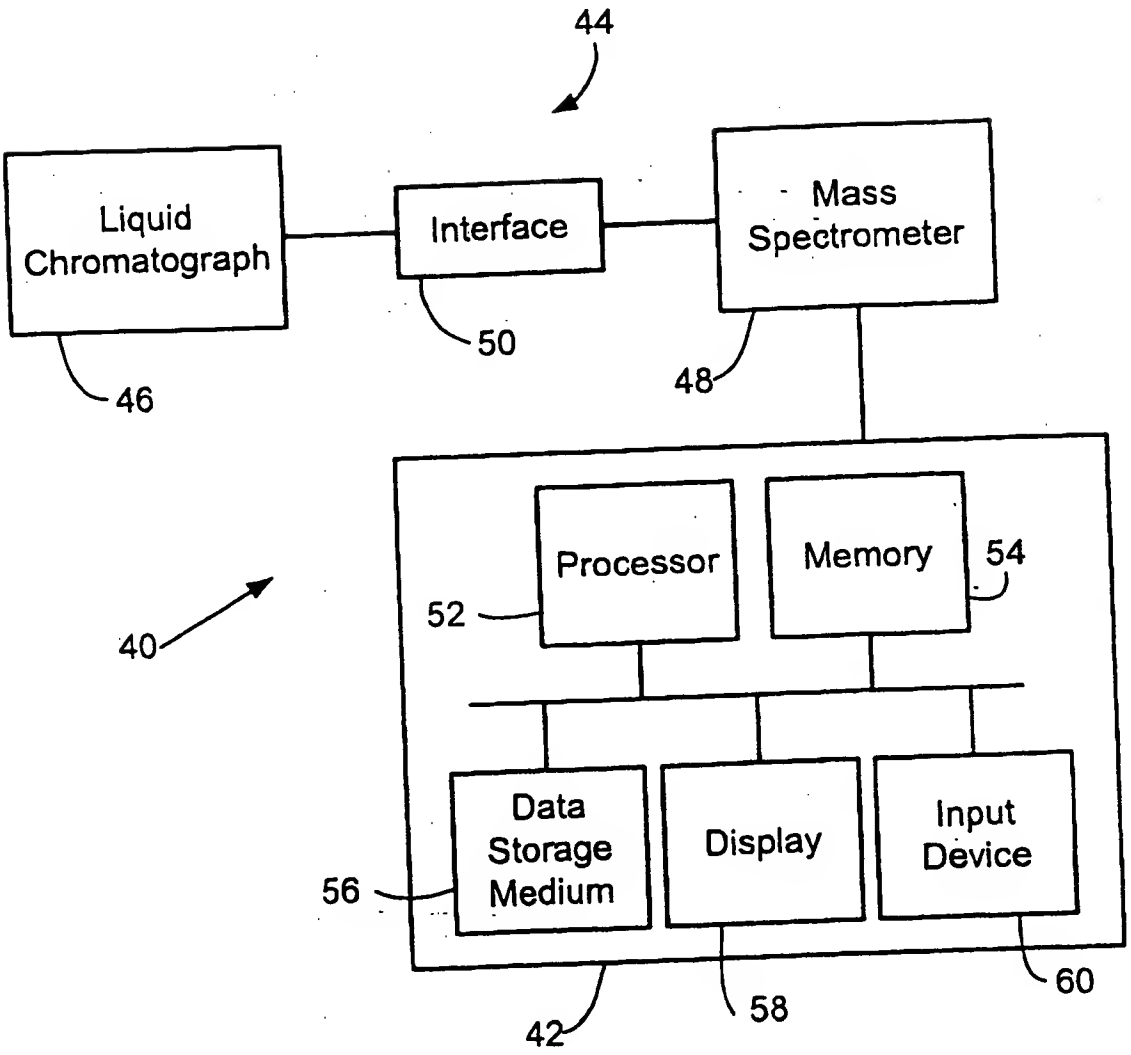

44
Liquid Chromatograph
Interface
Mass Spectrometer
46
50
48
40
Processor
Memory
52
54
Data Storage Medium
Display
Input Device
56
42
58
60

FIG. 6

SWANSON AND BRATSCHUN, L.L.C.
1745 SHEA CENTER DRIVE, SUITE 330
HIGHLANDS RANCH, CO 80129
PH. (303) 268-0066

U.S. BANK, N.A.
23-2/1020

10725

8/22/2001

PAY TO THE ORDER OF Commissioner of Patents and Trademarks $ **75.00

Seventy-Five and 00/100********** DOLLARS

Commissioner of Patents and Trademarks

MEMO SURR.64/PR

⑈010725⑈ ⑆102000021⑆ 1036565962 46⑈

DATE: August 24, 2001

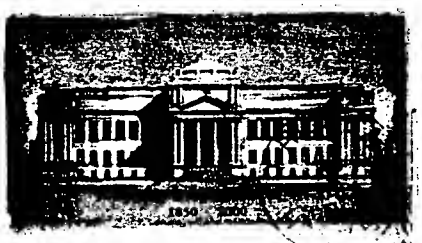
APPLICANT: HASTINGS *ET AL.*

SERIAL NO.: 60/_____

FOR: NONLINEAR FILTER FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA

RECEIPT IS HEREBY ACKNOWLEDGED OF: PROVISIONAL APPLICATION COVER SHEET; APPLICATION AS ENTITLED ABOVE (SPECIFICATION, 14 PGS.; FIGURES 1-6, 8 PAGES; TOTAL OF 22 PAGES.); CHECK IN THE AMOUNT OF $75.00 (S&B # 10725)

Docket No. SURR.64/PR
Express Mail No: EL 758770491 US



UNIVERSITY OF UTAH 20 USA

SWANSON & BRATSCHUN, L.L.C.
1745 SHEA CENTER DRIVE SUITE 330
HIGHLANDS RANCH, CO 80129

PA 1125407

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

February 10, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: *60/232,273*
FILING DATE: *September 13, 2000*

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

P. SWAIN
Certifying Officer

EXPRESS MAIL
UNITED STATES POSTAL SERVICE®

POST OFFICE TO ADDRESSEE

ORIGIN (POSTAL USE ONLY)

PO ZIP Code	Day of Delivery	Flat Rate Envelope
	☐ Next ☐ Second	☐
Date In		Postage
Mo Day Year	☐ 12 Noon ☐ 3 PM	$
Time In	Military	Return Receipt Fee
☐ AM ☐ PM	☐ 2nd Day ☐ 3rd Day	
Weight	Int'l Alpha Country Code	COD Fee / Insurance Fee
lbs. ozs.		
No Delivery	Acceptance Clerk Initials	Total Postage & Fees
☐ Weekend ☐ Holiday		$

DELIVERY (POSTAL USE ONLY)

Delivery Attempt	Time	Employee Signature
Mo. Day	☐ AM ☐ PM	
Delivery Attempt	Time	Employee Signature
Mo. Day	☐ AM ☐ PM	
Delivery Date	Time	Employee Signature
Mo. Day	☐ AM ☐ PM	

☐ **WAIVER OF SIGNATURE** (Domestic Only): **Additional merchandise insurance is void if waiver of signature is requested.** I wish delivery to be made without obtaining signature of addressee or addressee's agent (if delivery employee judges that article can be left in secure location) and I authorize that delivery employee's signature constitutes valid proof of delivery.

NO DELIVERY ☐ Weekend ☐ Holiday

Customer Signature

CUSTOMER USE ONLY

METHOD OF PAYMENT:

Express Mail Corporate Acct. No: X801638

Federal Agency Acct. No. or Postal Service Acct. No.:

FROM: (PLEASE PRINT) PHONE (303 268 0066

SWANSON & BRATSCHUN LLC
1745 SHEA CENTER DR STE 330
LITTLETON CO 80129-1540
USA

SURR.64 | PR

TO: (PLEASE PRINT) PHONE ( )

ASSISTANT COMMISSIONER
FOR PATENTS
WASHINGTON DC 20231-9999

BOX PROVISIONAL

PRESS HARD.
You are making 3 copies.

FOR PICKUP OR TRACKING CALL 1-800-222-1811 **www.usps.com** EMS

Mailing Label
Label 11-F August 2000

F02
T22

35/ 10